CLASSIFICAÇÃO NEURAL DE SINAIS DE SONAR PASSIVO

João Baptista de Oliveira e Souza Filho

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. Adrião Duarte Dória Neto, Dr.

Prof. Basílio de Bragança Pereira, Ph.D.

Prof. Carlos Eduardo Parente Ribeiro, D.Sc.

Prof. Luis Pereira Calôba, Dr.Ing.

Prof. Marley Maria Bernardes Rebuzzi Vellasco, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

JULHO DE 2007

OLIVEIRA E SOUZA, JOÃO BAPTISTA FILHO

Classificação Neural de Sinais de Sonar Passivo [Rio de Janeiro] 2007

XXIII, 310 pp 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia Elétrica, 2007)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1.Sonar Passivo 2.Redes Neurais Artificiais 3.Classificação 4.Extração de Características 5.Processamento de Sinais de Sonar 6.Análise de componentes principais (PCA) 7.Análise Estatística 8.Classificadores Modulares

I.COPPE/UFRJ II.Título (série)

Agradecimentos

Em primeiro lugar, a Deus, razão e causa de tudo que existe.

A minha primeira família terrena, nas figuras de meus queridos e saudosos pais e de minha tão companheira e amiga irmã. A eles, o agradecimento pelo sacrifício, renuncia e auxílio prestado nos mais variados campos de minha vida. Em especial, a gratidão pelo exemplo, motivação, e inspiração a eles devida, qualidades necessárias ao trilhar este caminho, que, ainda que tão significativo, é o primeiro passo de longa jornada.

A minha amada esposa Érika e, ao não menos amado, pequeno Lucas, pelo estímulo, amor, carinho e compreensão a mim dedicados, em especial, nos necessários períodos de ausência física. Destaque ao tão importante auxílio técnico prestado pelo pimpolho Lucas, nos seus primeiros meses de vida, ao analisar tão detalhamente o texto, rindo, amassando e comendo partes importantes da tese nas suas primeiras versões.

Ao meu orientador, Professor Seixas, pelas relevantes e extensas discussões críticas que contribuiram, de forma fundamental, na elaboração deste trabalho.

Ao laboratório de Processamento de Sinais (LPS) pela infra-estrutura disponibilizada, ao IPqM pelo fornecimento do banco de dados, à CAPES, ao CNPq e à FAPERJ pelo apoio financeiro fornecido.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

CLASSIFICAÇÃO NEURAL DE SINAIS DE SONAR PASSIVO

João Baptista de Oliveira e Souza Filho

Julho/ 2007

Orientador: José Manoel de Seixas

Programa: Engenharia Elétrica

Um sistema para a classificação do ruído irradiado por navios é uma ferramenta útil para os operadores de sonar e reduz a complexidade da tarefa de decisão. Este trabalho dedica-se à classificação neural de sinais de sonar passivo pré-processados, utilizando dados experimentais constituídos pelo ruído irradiado por 28 navios de 8 classes distintas. A análise de componentes principais (PCA) é utilizada para a extração de características do sinal e novos algoritmos de extração PCA, com operação em tempo real, eficazes e acurados são propostos. O estudo das restrições estatísticas no conjunto de dados é também desenvolvido. Para a tarefa de classificação, uma arquitetura baseada em sistemas especialistas é proposta, a qual utiliza uma rede neural *feedfoward* associada à identificação de cada classe. Esta arquitetura escalável mostra-se capaz de permitir a inclusão de novas classes, o que é tipicamente requerido em situações práticas. Uma eficiência média para a detecção das classes superior a 88,7% é obtida.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

NEURAL CLASSIFICATION OF PASSIVE SONAR SIGNALS

João Baptista de Oliveira e Souza Filho

July/ 2007

Advisor: José Manoel de Seixas

Department: Electrical Engineering

A system for automatic classification of the noise radiated by ships is an useful tool for sonar operators and reduces the complexity of the decision task. This work covers the neural classification of pre-processed passive sonar signals using experimental data from noise radiated by 28 ships belonging to 8 different classes. Principal component analysis (PCA) is used for signal feature extraction and new on-line, cost-effective and accurate algorithms for PCA extraction are proposed. A study of the observed statistical restrictions in dataset is also developed. For classification task, an expert network approach is proposed using a feedforward neural network assigned to each signal class. This scalable architecture is shown to allow the inclusion of new classes which is tipically required in pratical situations. An average efficiency for class detection better than 88.7% was obtained.

Conteúdo

Lista de Figuras

Lista de Tabelas

Siglas

AKC	*Adaptive Kernel Classifier*
ART	*Adaptive Resonance Theory*
APEX	*Adaptive Principal Components Extraction*
CNNE	*Construtive Neural Network Ensamble*
DNC	*Dynamic Node Creation*
EM	*Expectation-Maximization*
FLN	*Functional Link Network*
GHA	*Generalized Hebbian Algorithm*
KNN	*K-Nearest Neighbour*
LOFAR	*LOw Frequency Analysis and Recording*
LMSER	*Least Mean Square Error Reconstruction*
LPC	*Linear Predictive Coding*
LOO	*Leave-one-out*
LVQ	*Learning Vector Quantization*
MLP	*MultiLayer Perceptron*
MSE	*Mean Square Error*
NIC	*Novel Information Criteria*
PAST	*Projection Approximation Subspace Tracking*
PAW	*Physics Analysis Workstation*
PCA	*Principal Component Analysis*
PCD	*Principal Component for Discrimination*
PLA	*Polygonal Lyne Algorithm*
PSA	*Principal Subspace Analysis*
RCE	*Redundant Classification Environment*
RLS	*Recursive Least Square*

ROC	*Receiver Operating Characteristic*
RPROP	*Resilient PROPagation*
RKC	*Rapid Kernel Classifier*
SGA	*Stochastic Gradient Ascent*
SNC	*Sequencial Network Construction*
SNLA	*Subspace Network Learning Algorithm*
SOM	*Self-Organizing Maps*
STFT	*Short Time Fourier Transform*
TPSW	*Two-Pass Split Window*
WNIC	*Weighted Novel Information Criteria*
WSA	*Weighted Subspace Algorithm*

Parte I

Introdução

Capítulo 1

Contexto

O significativo progresso técnico-científico nas últimas décadas, em especial nos campos da física atômica, química, microeletrônica, telecomunicações e computação, viabilizou o desenvolvimento de sistemas complexos em várias áreas, tais como a Medicina e a Engenharia. Neste avanço tecnológico, número crescente de processos e sistemas tem sido instrumentados, notadamente após o advento da tecnologia digital, a qual viabilizou aplicações consideradas como de ficção científica apenas alguns anos atrás.

Os sistemas atuais, na sua maior parte baseados no processamento digital de informações, exploram técnicas sofisticadas para o tratamento dos sinais, em especial visando a compactação dos dados e/ou a extração de informação relevante, atuando na solução de problemas complexos, sujeitos a extenso número de variáveis. Número crescente de aplicações utilizam sistemas de inspiração biológica, sendo baseado em redes neurais artificiais, notadamente, quando envolvem o reconhecimento de padrões. Para aplicações complexas, é ainda freqüente a utilização de múltiplos agentes, numa estratégia de dividir-e-conquistar, onde múltiplos subsistemas especializados são dedicados à solução de uma mesma tarefa.

Os problemas envolvidos no ambiente de sonar são sabidamente complexos, e sua solução demanda, usualmente, o processamento digital em tempo-real de volume expressivo de informações, considerando técnicas complexas, de custo computacional elevado. Neste contexto, é de especial interesse a aplicação de técnicas de compactação e de sistemas neurais especializados. Cabe observar que graças ao progresso tecnológico, importantes ferramentas para a orientação e a defesa de embarcações

baseadas em sistemas de sonar puderam ser desenvolvidas.

1.1 Motivação

No meio aquático, os submarinos possuem um papel tático bastante importante na vigilância, a qual pode buscar a detecção de ameaças ou a identificação de objetos de interesse. Como o som, em meios submersos, é a única forma de energia que se propaga, de forma eficiente, a grandes distâncias, o sistema de sonar é o principal instrumento utilizado nos submarinos para esta vigilância.

Sistemas de sonar podem ser ativos ou passivos. Nos sonares ativos, há a emissão de ondas acústicas e a análise do seu retorno. Para os passivos, realiza apenas a captação e análise do ruído aquático. Como a emissão de ondas acústicas poderia denunciar sua presença e localização, os submarinos, em situação de vigilância, utilizam sonares passivos, mantendo-se ocultos, a fim de não comprometer este processo ou tornar-se demasiado vulneráveis.

Embarcações no entorno de um submarino são referidas como contatos. A detecção e classificação dos contatos é de especial importância para a operação e defesa dos submarinos. Cada contato produz um conjunto de ruídos característicos, referidos como assinatura acústica, relacionados a características físicas e de maquinário em operação no seu interior. A classificação dos contatos é normalmente realizada por operadores especialmente treinados para esta tarefa, sendo baseada na sonoridade dos sinais capturados e por informações adicionais, tais como seu conteúdo de freqüência. Trata-se de uma tarefa especialmente difícil e estressante, tendo em vista a complexidade dos sinais envolvidos, o número expressivo de classes existentes e as múltiplas condições operativas e de cenário aquático possíveis.

Prover um sistema de apoio ao operador de sonar que identifique, de forma automática, a qual classe o contato pertence, assim como forneça eventuais análises e informações complementares, é de extrema relevância. Adoção deste sistema em cenários de operação militar pode resultar na redução da carga de trabalho e do estresse emocional e físico, assim como numa maior rapidez e confiabilidade na tomada de decisões.

1.2 Classificação automática de contatos

Em razão da complexidade do problema e de requisitos quanto ao desempenho e à confiabilidade, a constituição de um sistema de classificação automática para o ambiente de sonar não é uma tarefa trivial. Um requisito fundamental é dispor de uma base de dados que caracterize, apropriadamente, o problema, o que exige, normalmente, diferentes aquisições do ruído irradiado para várias classes e navios, segundo diferentes situações operativas e ambientais, em condições similares àquelas a serem defrontadas pelo sistema em cenários reais de operação.

Tendo em vista a multiplicidade dos cenários operativos possíveis e restrições quanto ao custo, tempo de aquisição e o número de navios envolvidos, é provável a existência de restrições quanto à caracterização das classes. Característica fundamental ao sistema classificador é, ainda que desenvolvido sobre dados que possuam estas restrições, em especial com respeito a cenários operativos reais, apresente uma capacidade de generalização adequada à sua aplicação como sistema de apoio, visto que é provável deparar-se com sinais de características estatísticas distintas àquelas consideradas no seu desenvolvimento.

Deste modo, o sistema de classificação automática deve considerar, em razão da complexidade do problema e da alta-dimensionalidade dos dados envolvidos, sistemas híbridos de classificação, onde técnicas de compactação, constituídas por cadeias de pré-processamento baseadas em conhecimento especialista, e sistemas de extração adaptativa de características lineares e não-lineares são combinadas com classificadores neurais, hábeis no reconhecimento de padrões em espaços de dados de dimensão elevada, mesmo em problemas com severas restrições quanto sua caracterização estatística. A adoção de classificadores neurais modulares na constituição destes sistemas híbridos é atrativa à operação em tempo real do sistema, permitindo, também, uma identificação e incorporação facilitada de novos cenários, em especial novas classes.

1.3 Objetivos da tese

Um dos objetivos deste trabalho é contemplar alguns dos estágios necessários à produção de um sistema de classificação automática de contatos, na qualidade de

um equipamento para operação embarcada, desenvolvido com tecnologia própria, a ser incorporado ao sistema de sonar de submarinos da Marinha Brasileira. Este trabalho é decorrente de uma colaboração estabelecida entre o Instituto de Pesquisas da Marinha (IPqM) e a Universidade Federal do Rio de Janeiro (UFRJ), e representa uma continuidade de trabalhos anteriores, em especial, a referência [2], que também utilizou dados de raia acústica, para um conjunto mais restrito de classes e contatos, enfocando, no entanto, a constituição do pré-processamento e a extração de características, através de técnicas lineares e não-lineares, adequados a uma apropriada classificação destes sinais.

Os esforços deste trabalho se concentraram em duas áreas principais: a extração de características e o projeto e avaliação de classificadores neurais, ambos realizados com sinais de raia acústica, pré-processados por uma cadeia similar à proposta em [2]. Segundo este pré-processamento, uma classificação deve ser produzida a cada $0,2$ s, tempo que é compatível, para as técnicas utilizadas neste trabalho, com a implementação do sistema classificador para a operação em tempo real [3], em especial quando realizada através de processadores digitais de sinais modernos.

Quanto à extração de características, ênfase especial é dedicada à proposição de um sistema extrator adaptável, útil à incorporação, em operação, de novos cenários ao sistema classificador. Para este sistema, é proposta a utilização das direções fornecidas pela análise de componentes principais (PCA) [4], sendo realizada uma extensa discussão sobre algoritmos adaptativos para a extração das componentes, a qual contempla uma detalhada revisão da literatura, assim como a proposição de algoritmos rápidos, acurados e capazes do operar em tempo-real, ambos avaliados quanto à acuidade e o custo computacional na extração de componentes do conjunto do sonar.

Quanto à classificação, é abordado o projeto de classificadores neurais em cenários com requisitos severos quanto à confiabilidade e à capacidade de generalização, assim como sujeitos a prováveis restrições quanto à caracterização estatística das classes utilizadas para o seu treinamento. Detalhada discussão de índices e técnicas estatísticas para a avaliação da capacidade de generalização de classificadores neurais é desenvolvida. Realiza-se uma caracterização do banco de dados utilizado nesta tese, através de classificadores MLP totalmente conectados [4], sendo identi-

ficadas as classes de classificação mais crítica. Através desta caracterização, foram identificadas restrições quanto à caracterização das condições operativas para a base de dados em estudo.

Uma vez que pelas restrições identificadas quanto à caracterização das classes, o desempenho quanto à generalização mostrou-se sensível à escolha dos conjuntos utilizados no projeto e avaliação do classificador, é proposta que esta escolha seja baseada em grupos identificados por análise de agrupamentos. Para sua produção são apresentadas técnicas aplicáveis ao problema de sonar, sendo discutida a escolha dos parâmetros utilizados pelos algoritmos, assim como analisado o impacto desta escolha no desempenho do classificador. Como resultado são produzidos conjuntos representativos do problema segundo dois enfoques: um primeiro, em que o sistema é avaliado segundo as mesmas condições operativas utilizadas no seu treinamento (projeto e avaliação baseados em espectros); e uma segunda, onde o projeto e a avaliação podem considerar diferentes condições operativas (projeto e avaliação baseados em corridas).

É proposto um sistema de classificação baseado em classificadores classe-especialistas, isto é, classificadores formados por diferentes módulos, cada qual especializado no reconhecimento de uma classe. Entre as técnicas discutidas para a constituição deste sistema, tem-se: os filtros casados, as curvas principais e as redes modulares classe-especialistas, a última discutida extensivamente. Entre atrativos, as redes modulares são aplicáveis à construção de classificadores para número expressivo de classes e ambientes de alta-dimensionalidade e complexidade, e permitem a detecção, a inclusão de novas classes e a atualização parcial do sistema, características nitidamente desejáveis no ambiente de sonar. Para esta proposta, são discutidos o treinamento e o dimensionamento dos especialistas, assim como critérios para a detecção e a inclusão de novas classes.

Por basear-se em aquisições realizadas em raia acústica, a qual possui baixa profundidade, e representa um ambiente de medição controlado, onde o sinal adquirido pertence a um único contato e não há contaminação do ruído produzido pelo próprio submarino (ruído próprio), cabe aos futuros trabalhos explorar dados adquiridos em águas profundas através de operações militares envolvendo o submarino e os diferentes contatos. Neste caso, entre aspectos relevantes à pesquisa a ser

desenvolvida, tem-se a separação de contatos simultâneos e a eliminação do ruído próprio.

1.4 Contribuções da tese

As principais contribuições deste trabalho residem no desenvolvimento de um sistema de apoio à classificação de contatos utilizando uma base de dados disponibilizada pelo IPqM, e na proposição de ferramentas úteis ao projeto de classificadores neste ambiente complexo, sujeito a eventos de elevada dimensionalidade e a prováveis restrições quanto à caracterização estatística das classes, sendo resumidas a seguir:

- Quanto à extração de características:
  - Discussão de técnicas adaptativas aplicáveis à extração PCA, a qual contempla uma revisão da literatura e a proposição de novos algoritmos. Nesta revisão, os diferentes algoritmos são apresentados segundo uma notação unificada, e são discutidas semelhanças quanto às funções objetivo e às arquiteturas utilizadas para a extração, informações relevantes à seleção de um algoritmo que vise uma aplicação particular. Trabalhos anteriores identificaram a PCA como uma metodologia de compactação útil à classificação dos sinais de sonar [2, 5], porém não discutem técnicas para sua obtenção. É proposta uma nova arquitetura neural de extração, para a qual são derivados algoritmos acurados e de custo computacional reduzido, atrativos ao problema de sonar.
- Quanto à classificação:
  - Diferentes técnicas são apresentadas para a construção de classificadores classe-escaláveis. Ênfase é dedicada às redes modulares classe-especialistas, para a qual o impacto do treinamento e do dimensionamento dos classificadores na eficiência de generalização é discutido, aspecto não verificado na literatura. Critérios de integração dos especialistas altenativos ao máximo, comumente explorado, são analisados. É ainda discutida a detecção

e inclusão de novas classes ao sistema, visto que são raros os trabalhos que abordam este tópico, em especial em aplicações de sonar passivo.

- Com respeito a ferramentas de projeto:
  - São propostos mecanismos para a caracterização estatística dos dados e realizada uma identificação das classes de classificação mais crítica, que visa orientar novas aquisições de dados e a tomada de decisão pelo operador.
  - É discutida a seleção de conjuntos estatisticamente representativos da base disponível para o projeto e avaliação do classificador, realizados através de agrupamentos de espectros e corridas. São também propostas técnicas de seleção dos parâmetros envolvidos na produção destes agrupamentos que utilizam o próprio classificador como figura de mérito.

1.5 Estrutura do documento

O documento foi dividido em três partes, visando uma melhor organização e clareza.

Na parte I, é realizada uma introdução ao problema, sendo constituída pelo presente capítulo, e pelo Capítulo 2, onde é realizada uma breve introdução ao problema e aos sistemas de sonar passivo, sendo discutidas algumas características dos sinais envolvidos e a complexidade da classificação dos contatos. Em seguida, é realizada uma revisão bibliográfica de trabalhos publicados sobre a classificação de contatos, em especial, utilizando redes neurais. Por fim, são discutidas características do banco de dados utilizado no desenvolvimento desta tese e realizadas considerações sobre o sistema de classificação discutido neste trabalho.

A parte II é constituída pelo Capítulo 3, abordando o pré-processamento e a compactação dos sinais, em especial a extração adaptativa de componentes principais, para qual é realizada uma extensa revisão da literatura, onde os diferentes algoritmos são agrupados, didaticamente, por semelhança estrutural ou do processo de derivação. É proposta uma arquitetura neural para a extração das componentes, a qual é explorada para a produção de algoritmos de operação *on-line* e *off-line*.

Realiza-se, também, uma comparação deste algoritmos com técnicas consagradas da literatura na extração de componentes para o conjunto de dados de sonar.

A parte III discute a produção de sistemas de classificação neural para o ambiente de sonar passivo, sendo constituída pelos Capítulos 4, 5 e 6.

No Capítulo 4 são discutidos aspectos que influenciam na capacidade de generalização de classificadores neurais, assim como índices e técnicas para sua avaliação. Em seguida, é realizada uma caracterização dos dados considerando o projeto e avaliação por espectros e corridas, e identificadas as classes sujeitas a restrições quanto à caracterização das condições operativas mais críticas.

O Capítulo 5 é dividido em duas partes principais: a seleção baseada em espectros e em corridas. Na primeira, é realizada uma introdução sobre a análise de agrupamentos, em especial através das técnicas de agrupamento seqüencial e hierárquico, assim como a seleção dos parâmetros envolvidos, a qual é realizada através da proposição de alguns critérios e pela avaliação de índices estatísticos da literatura. Na segunda, é discutido um critério que identifica similaridades entre as corridas, realizada através da definição de um vetor representativo associado a cada uma, e posterior produção e análise de um agrupamento baseado nestes vetores.

No Capítulo 6 é discutida a constituição do sistema de classificação de contatos através de múltiplos classificadores, com ênfase à técnica de classificadores modulares classe-especialistas. Para esta técnica, são apresentados critérios para o treinamento, definição da topologia e integração dos classificadores. O capítulo é encerrado abordando a detecção e inclusão de novas classes a este sistema.

Por fim, finalizando este trabalho, o Capítulo 7 apresenta as conclusões e descreve possíveis trabalhos futuros.

Capítulo 2

Sonar Passivo

Os dois principais sentidos utilizados pelos seres humanos na captação de informações do meio que os rodeiam são a visão e a audição. Para ambos sentidos, o tipo de energia radiante envolvida, ou seja, a luz e o som, propaga-se facilmente pelo ar. Em ambientes aquáticos, no entanto, o som é a única forma de energia que se propaga de forma eficiente por longas distâncias. Assim, o conhecimento e a exploração do meio aquático, realizado tanto pelos animais quanto pelos seres humanos, é obtido pela captação e processamento de energia acústica. O conhecimento da utilidade dos sons aquáticos é antigo, remontando, pelo menos, ao século 15, quando Leonardo D'Vinci fez a afirmação de que a escuta de um tubo introduzido na água permite a audição de navios situados a uma grande distância, idéia que é a base dos sistemas de sonar atuais [6].

Uma das mais importantes tarefas realizadas pelas forças armadas é a vigilância, a qual pode visar a detecção de ameaças ou de objetos de interesse. A rapidez e a confiabilidade desta identificação são requisitos importantes, a fim de que medidas apropriadas sejam tomadas em tempo hábil. Para realizar esta tarefa, variados equipamentos são utilizados, e no meio aquático, os submarinos possuem um papel tático bastante importante, explorando, para o desempenho desta função, os sistemas de sonar.

Existem dois tipos básicos de sistema de sonar: o ativo e o passivo. No sonar ativo, o sistema transmite uma energia acústica através da água, analisando seu retorno (eco), de forma análoga ao sistema de radar [7]. No sonar passivo, não há o envio de energia acústica, existe apenas a captação e análise do ruído aquático, o que

permite ao observador se manter silencioso (oculto), na escuta dos demais [8]. Para a operação e defesa de submarinos, dada sua função estratégica, é utilizado o sistema de sonar passivo, visto que a emissão de ondas acústicas necessária aos sonares ativos poderia denunciar sua presença, tornando-os vuneráveis e comprometendo o processo de vigilância.

A classificação de contatos, ou seja, a identificação de possíveis ameaças no entorno de um submarino é realizada, normalmente, pelo operador de sonar. Cada contato produz um ruído característico, referido como assinatura acústica, o qual está relacionado às suas características físicas, tais como o número de pás e hélices, o tipo de propulsão (diesel ou elétrica), assim como o tipo e a configuração das máquinas em operação no seu interior. As máquinas do contato, em especial as rotativas, tais como motores e geradores, produzem fundamentais e harmônicos característicos. Historicamente, os operadores de sonar realizavam esta operação através da audição do sinal captado pelo sistema [9]. Com o desenvolvimento da eletrônica, dos algoritmos de processamento de sinais [6] e dos computadores digitais, informações adicionais, tais como o conteúdo espectral do sinal, entre outras análises de apoio, são disponibilizadas ao operador, o que permitiu uma redução do tempo de reação e uma maior confiabilidade nas interpretações e na tomada de decisão [10].

A identificação dos contatos pelo ruído captado pelo sistema de sonar não é uma tarefa trivial. Embora haja bastante conhecimento na literatura sobre a natureza e a geração de sinais aquáticos, as características do ruído captado variam de acordo com o período do ano, localização, condições de propagação (a propagação em águas rasas difere daquela em águas profundas) e condições ambientais. Em geral, os ruídos de interesse (produzidos pelo contato) estão ainda imersos num ruído de fundo, o qual inclui os ruídos do próprio submarino, naturais (ondas, ventos na superfície, raios, turbulência, entre outros) e biológicos (baleias, golfinhos, etc.). Este rico conjunto ainda sofre múltiplas reflexões na superfície e no assoalho submarino, problema que é agravado em águas rasas [6]. Tratam-se, pois, de sinais altamente não-estacionários e impulsivos, que apresentam variações rápidas nas características espectrais, em termos de freqüência e tempo, e nas suas relações sinal-ruído, em virtude da multiplicidade de fontes e de caminhos percorridos até o observador [11].

Atualmente, a classificação de contatos é baseada na análise de sonoridade do

sinal e na realização da análise LOFAR (*LOw Frequency Analysis and Recording*). Nesta análise, é visualizado um gráfico (espectrograma), onde cada ponto corresponde ao valor da densidade espectral de potência [12] para uma dada freqüência (definida pelo eixo horizonal) e janela de tempo (correspondente ao eixo vertical) [13]. Navios, submarinos e torpedos constituem excelentes fontes de ruído no mar, uma vez que possuem máquinas rotacionais e alternativas para a propulsão, o controle e a habitabilidade [2]. Em [2], tem-se um descrição resumida das principais características de um sistema de sonar passivo e do ruído irradiado por embarcações. Na Figura 2.1, com fins ilustrativos, apresentamos um espectrograma de uma corrida típica para um dado navio.

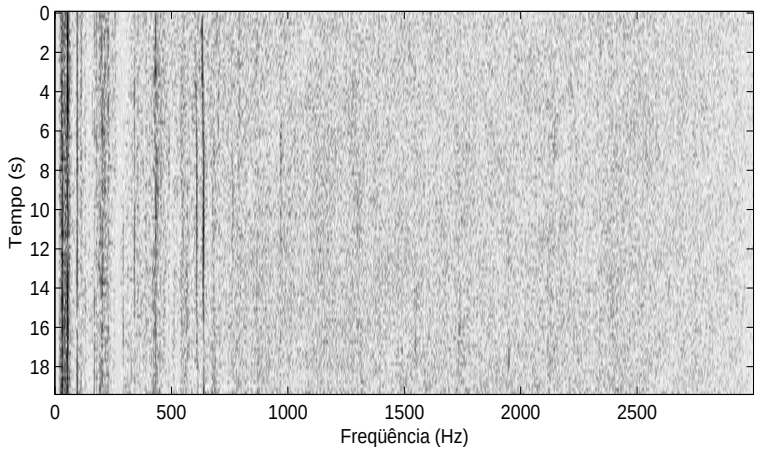


Figura 2.1: Espectrograma de uma corrida arbitrária.

2.1 Classificação automática de contatos

Tendo em vista a complexidade envolvida na classificação de um contato, prover um sistema automático para sua classificação é uma ferramenta bastante útil ao trabalho do operador de sonar, constituindo um importante equipamento de apoio à decisão. A adoção de sistemas de apoio à decisão normalmente reduz a carga de trabalho do operador, o qual pode se concentrar em contatos de maior relevância. De posse de um maior volume de informações sobre o contato, a rapidez e a confiabilidade do processo de tomada de decisões é elevada, o que é vital em

cenários de conflito. Sistemas automáticos podem ainda compensar deficiências do operador, em especial a fadiga, a qual gera um estado conhecido como decremento de vigilância [9].

O desenvolvimento de um sistema de classificação eficiente exige uma base de dados rica, que contemple o maior número de cenários possível, ou seja, diferentes aquisições do ruído irradiado para várias classes e navios, segundo diferentes condições operativas e ambientais, em condições o mais próximas do real possível. Entretanto, há uma série de restrições práticas relacionadas à produção deste conjunto. O número de classes e navios existentes é expressivo, e cada classe, freqüentemente, subdividide-se em várias subclasses, que podem possuir características de ruído irradiado distintas.

A aquisição do ruído irradiado por contatos é um processo caro e longo, devendo ser realizada, preferencialmente, em cenários controlados, o que exige, freqüentemente, o deslocamento da embarcação para uma localidade específica. É necessário ainda submeter as embarcações envolvidas a diferentes condições operativas, cuja gama de possibilidade pode ser bastante ampla. O número de classes e navios à disposição da marinha para estes ensaios é outra restrição importante. Significativo é o número de embarcações existentes e nem todos navios de interesse podem estar disponíveis, sendo possível que estejam alocados para outras missões em curso, ou mesmo, pertencer a outras marinhas. No último caso, a aquisição do ruído pode ter sido realizada em exercícios conjuntos com marinhas amigas, de forma oculta, o que restringe a quantidade de condições operativas disponíveis. É provável, portanto, a existência de restrições estatísticas na caracterização das diferentes classes, assim como é esperado que o sistema desenvolvido, quando em operação, se depare com novas condições operativas, navios ou mesmo classes de navios.

Em razão das características do problema, uma série de exigências são impostas ao sistema de classificação. A capacidade de generalização, isto é, do sistema operar corretamente em condições operativas distintas àquelas consideradas no seu projeto é fundamental. O sistema deve, portanto, mostrar-se insensível a suposições de projeto e a ambientes específicos de operação [14]. Como o sistema de classificação automática operará em tempo real, o custo computacional envolvido deverá ser considerado na escolha do conjunto de técnicas utilizadas, o que conduz, em razão

da alta-dimensionalidade dos dados, à utilização de técnicas de extração de características e de compactação dos dados. Eventualmente, implementações em *hardware*, através de processadores digitais de sinais ou mesmo lógica programável, podem ser necessárias [15]. Outros fatores relevantes nesta escolha são o grande número de classes existentes, assim como a necessária identificação e incorporação de novos cenários, em especial novas classes.

Classificadores automáticos exigem, normalmente, técnicas que identifiquem características dos dados que são relevantes à sua discriminação nas diferentes classes. Para o reconhecimento de assinaturas fixas em meio a ambientes estacionários, variadas técnicas eficientes foram desenvolvidas. Quando, no entanto, as assinaturas e o meio são variáveis, de forma conhecida ou não, as técnicas tradicionais tendem a não apresentar bons resultados. Uma alternativa interessante tanto ao pré-processamento quanto à classificação dos sinais de sonar passivo são as redes neurais artificiais [4], dada sua eficiência na extração de características relevantes, mesmo em ambientes de alta-dimensionalidade e complexidade. Tratam-se de sistemas adaptativos que aprendem através de exemplos [16], os quais não exigem uma modelagem prévia das características estatísticas dos sinais envolvidos e apresentam um bom compromisso em termos do tempo de treinamento, requisição de memória e complexidade de código [17], o que viabiliza sua implementação em tempo real. As redes neurais são ainda estruturas modulares, capazes de extrair ou modelar, adaptativamente, informações estatísticas de alta-ordem dos sinais, o que facilita a classificação de sinais não-gaussianos [18].

A classificação de sinais de sonar passivo em diferentes trabalhos da literatura usualmente explora a combinação de uma cadeia de pré-processamento com classificadores neurais [11, 19, 20, 21, 22]. Em linhas gerais, nas cadeias de pré-processamento é realizada uma estimação do conteúdo espectral do sinal na faixa de interesse, em geral inferior a 5 kHz, através de transformadas, tais como: a transformada rápida de Fourier (FFT) [23], a transformada de Fourier de tempo curto (STFT) [23], ou ainda, através dos coeficientes da expansão do sinal em *wavelets* [24]. Em alguns trabalhos [2, 9, 25, 26], após o pré-processamento, são utilizados extratores de características que visam reduzir a dimensão dos dados e enfatizar as características relevantes à classificação.

Este trabalho se concentrará em discutir aspectos relevantes à constituição de um sistema de classificação automática de contatos, a ser incorportado, como equipamento, aos sistemas de sonar de submarinos da Marinha brasileira. Em razão do número expressivo de classes envolvidas e da complexidade do ruído irradiado, este sistema é estruturado em três módulos, a saber: um sistema de pré-processamento, um extrator de características e um módulo classificador neural. Pelo primeiro, estima-se o conteúdo espectral na faixa de interesse ($0 \approx 3kHz$), eliminando o ruído de fundo do ambiente de medição e enfatizando os tons espectrais, que são relavantes à caracterização das classes. Quanto ao extrator de características, seu objetivo é prover uma estatística suficiente [27] ao classificador, e apresentar, juntamente com o classificador, a capacidade de adaptação, visando viabilizar a identificação e incorporação de novos cenários, em especial classes. Para o classificador é esperado um bom desempenho na identificação das classes, em especial, em cenários reais de operação, ainda que sua produção tenha utilizado um conjunto de dados com restrições quanto ao número de classes e as condições operativas disponíveis. Outra exigência, comum aos três módulos, é possuir um custo computacional compatível com a operação *on-line* do sistema.

A seguir será apresentada uma revisão da literatura sobre a classificação de sinais de sonar passivo, com ênfase às propostas baseadas em redes neurais. Em seguida, será apresentado o banco de dados utilizado para o desenvolvimento deste trabalho. Por fim, serão realizadas considerações sobre a avaliação do sistema de classificação analisado nesta tese.

2.1.1 Revisão da literatura

Grande parte dos trabalhos referentes à classificação de sinais de sonar são de divulgação restrita, em especial durante o período de guerra-fria [2]. Para as publicações irrestritas, a técnica de redes neurais artificiais é freqüentemente explorada, apresentando resultados significativos. Para boa parte dos trabalhos, tem-se sistemas híbridos de classificação, onde um sistema de pré-processamento ou extração de características é responsável pela extração de informação inteligente dos dados, fornecendo uma estatística suficiente para a operação do classificador neural.

Gorman e Sejnowski (1988) [19] avaliaram a aplicação de redes neurais para

a classificação de ecos de sonar ativo, buscando distinguir entre rochas e minas. Os sinais utilizados foram reais, sendo submetidos, a fim de alimentar as entradas de uma rede neural tipo MLP [28], à seguinte seqüência de pré-processamento : cálculo da *Short-Time Fourier Transform* (STFT) [23], extração da envoltória espectral e normalização na faixa de 0 a 1. Os autores analisaram ainda o impacto da seleção dos conjuntos de treinamento e teste, do dimensionamento da rede (número de camadas e de neurônios na camada escondida) e da inicialização dos pesos na eficiência de classificação e generalização da rede. Foi verificado que a seleção mais criteriosa dos conjuntos de treino e teste produziu impacto na capacidade de generalização do classificador, aumentando sua eficiência média e reduzindo a dispersão obtida nos vários ensaios realizados. Constatou-se também que o impacto da inicialização dos pesos não foi significativo, reduzindo-se conforme o aumento do número de neurônios da camada escondida, o qual também produziu um aumento da eficiência de classificação e generalização médias, acompanhadas de uma redução da dispersão de ambas, até um limite de 12 neurônios. Para esta configuração, as eficiências obtidas foram de 99,8% e 90,4% para os conjuntos de treino e teste, respectivamente.

Khotanzad et al. (1989) [29] propuseram um sistema para a classificação de sinais de sonar passivo utilizando tanto o pré-processamento quanto a classificação baseados em técnicas neurais. O pré-processamento utilizado visou eliminar o ruído de fundo, reconstruindo o sinal recebido. Para esta tarefa, foi utilizada uma memória auto-associativa baseada na topologia de Hopfield [30] submetida, no entanto, a um algoritmo distinto de aprendizado e operação. A classificação é baseada em redes MLP, utilizando imagens simuladas de espectrograma de 2 classes de interesse (alvo e não-alvo). Foram reportadas eficiências de 97% para a classe alvo e de 100% para a classe não-alvo. Van-Houtte et al. (1991) [31], baseando-se no trabalho anterior, propuseram uma estrutura semelhante de classificação, substituindo, no entanto, o classificador MLP por uma memória associativa bidirecional [32]. O artigo reportou que esta técnica produziu classificadores melhores que os baseados em redes MLP, especialmente para relações sinal-ruído mais críticas, embora uma comparação qualitativa das técnicas não seja realizada.

Moore et at. (1991) [33] analisaram uma nova arquitetura neural conhecida como *Integrator Gateway Network* para a classificação de múltiplos ecos produzidos

na atuação do sistema de sonar biológico dos golfinhos. Como entradas da rede foram utilizados vetores espectrais dos ecos e o início de cada varredura. Foi mostrado que a combinação da informação dos múltiplos ecos pelo método produz resultados melhores de classificação, tendo sida atinginda uma faixa de eficiência de 90 a 93%, do que uma rede MLP de topologia similar, a qual atingiu apenas 63 % de eficiência.

Russo (1991) [25] estudou a classificação de sete tipos diferentes de sinais de sonar passivo através de sua forma característica, visualizada pelo espectrograma. Foi utilizado um sistema de pré-processamento baseado em técnicas morfológicas para a extração do contorno das raias espectrais, de forma a produzir um código chave [34], que define a entrada do classificador. O classificador foi formado por um conjunto de três redes neurais, todas elas treinadas separadamente, utilizando, no entanto, o mesmo conjunto de treinamento. Cada classificador possuía 7 saídas, que foram treinadas de forma que um único neurônio estivesse ativo por classe. Os eventos foram atribuídos às classes somente quando a saída correspondente apresentou valor superior a um limiar específico; caso contrário, o evento foi identificado como de classe desconhecida. Para a classificação final foi estabelecido um critério de votação, pelo qual o evento era identificado como pertencente a uma dada classe se, e somente se, pelo menos 2 classificadores apresentassem saídas concordantes. Utilizando um conjunto de dados predominantemente constituído por eventos reais, foi obtida uma eficiência de 93% de deteção, 5% de erros de classificação, sendo 2% dos eventos classificados como desconhecidos. Mostrou-se ainda que a combinação dos classificadores permitiu um aumento de eficiência de detecção em, pelo menos, 2 pontos percentuais.

Calsselman et al. (1991) [20] avaliaram a detecção e a classificação de sinais de sonar passivo utilizando o banco de dados *DARPA Standart Set I* [35], o qual é formado por quatro classes que contêm transientes de curta duração, e duas classes que contêm tons. O sistema proposto possuía uma dupla cadeia de pré-processamento e classificação, uma voltada ao tratamento de transientes, e a outra responsável pelos tons. Basicamente, ambas cadeias realizam uma estimação da densidade espectral de potência do sinal. A classificação dos transientes utilizou uma rede neural MLP, enquanto a de tons, um algoritmo baseado na comparação da energia de janelas espectrais com limiares. Para o conjunto de teste, uma eficiência

mínima de 79%, e máxima de 100%, foi obtida para cada classe, o que resulta numa eficiência de classificação média de 95,4%. Utilizando o mesmo banco de dados, Ghost et al. (1992) [11] discutiram a classificação neural de transientes de sonar passivo utilizando um pré-processamento baseado na transformada *wavelet* [24]. Foram apresentados classificadores baseados em núcleo (*Kernel*) [36], sob a alegação de possuírem maior robustez ao ruído e à dimensionalidade dos dados. Afirmou-se ainda que a redução da dimensionalidade dos dados, realizada através da extração de componentes principais [37], segundo o método de Sanger [38], assim como a utilização de mapas auto-organizáveis de Kohonen (SOM - *Self-organazing maps*) [39], não apresentaram bons resultados. Foram propostos três algoritmos para o treinamento do classificador: o *AKC - Adaptive Kernel Classifier*, cujo ajuste do centro dos núcleos utiliza a regra delta [4]; o RKC (*Rapid Kernel Classifier*), com ajuste dos centros pela técnica de LVQ [39]; e, por fim, o método RKCEB, o qual é similar à técnica RKC, utilizando, no entanto, núcleos elípticos. Foi realizada uma análise comparativa dos métodos propostos com outras metodologias consagradas, tais como: o método dos vizinhos mais próximos (KNN) [39], as redes de base radial [40], o método LVQ-2 [39] e as redes PI-SIGMA [41]. Os melhores resultados foram obtidos para o método AKC, atingindo uma eficiência de 100% para os sinais de teste. O desempenho destas técnicas foi ainda reavaliado para um conjunto de teste mais crítico, que busca simular condições de oceano mais realísticas. Para este conjunto, a eficiência se situou na faixa de $21,1\%$ a $31,6\%$, mostrando uma demasiada especialização dos classificadores no conjunto de treinamento. Por fim, foi proposta a combinação dos classificadores treinados segundo dois métodos: o integrador baseado em entropia e a combinação heurística de fatores de confidência. O segundo método apresentou melhores resultados, sendo atingida uma eficiência de 47 % para o conjunto mais realístico.

Weber e Krüger (1993) [21] propuseram a detecção e a classificação de tons de espectrograma utilizando uma rede neural de realce de contono (*contour enhancing* neural network) [42] e redes MLP. Para o conjunto de sinais simulado, uma eficiência de classificação de 95 % foi atingida, considerando sinais com relação sinal-ruído de -17dB.

Hemminger e Pao (1994) [9] utilizaram o conjunto DARPA para a detec-

ção e a classificação de transientes. O sistema proposto foi constituído por um sistema de pré-processamento, uma rede auto-organizável [4] e um conjunto de redes classificatórias baseadas na arquitetura *Functional-Link Net* (FLN) [43]. O pré-processamento consistiu no janelamento do sinal em seqüências de, aproximadamente, 10 ms (256 amostras), com superposição (*overlap* [44]) de 75%. Cada janela foi submetida a uma STFT e o espectro resultante foi dividido em sete segmentos, produzindo trechos de 64 amostras (50 % de superposição). Cada trecho sofreu ainda um janelamento de Hamming [44] e o cálculo da transformada de Fourier, visando a estimação da densidade espectral de potência, método que é conhecido como periodograma de Welch [45]. As redes auto-organizáveis utilizadas possuiam um raio de vigilância fixo e utilizaram a distância de Hausdorf [46] como métrica, sendo treinadas com dados provenientes de todas as classes. Para a classificação foram identificados, classe-a-classe, os protótipos mais representativos do mapa SOM, isto é, os neurônios mais freqüentemente excitados para cada um dos trechos de uma dada janela (7 no total). Uma rede FLN foi treinada para cada classe, utilizando as distâncias normalizadas do espectro de entrada aos protótipos característicos de cada classe. Para a combinação dos resultados dos classificadores foi utilizado o critério de máximo, isto é, o classificador vencedor (aquele que apresentou maior valor de saída) definiu a classe à qual o evento pertence. Como resultados, a eficiência por classe obtida para o conjunto de teste foi de, no mínimo, 88% e, no máximo, 100%, totalizando uma eficiência média de 96,8%.

Ward e Stevenson (2000) [22] avaliaram o uso da rede de resposta finita ao impulso (FIRNN - *Finite Impulse Response Neural Network*) [47] para a detecção contínua e a classificação de transientes de sonar passivo. O banco de dados utilizado consistiu em 15 classes de sinais distintas, adquiridas pelo orgão *Defense Research Establishment Atlantic* (DREA) [48], na bacia de Bedford, Canadá. O sistema proposto foi formado por uma cadeia de pré-processamento e por uma rede classificatória. O pré-processamento realizou a re-amostragem [44] do sinal para 12,5 kHz (amostragem original: 25 kHz), dado que a informação relevante à classificação se encontrava na faixa de 0 a 5 kHz. Em seguida, o sinal foi submetido a uma STFT com janela de Hamming de 256 pontos, resultando numa resolução em freqüência de, aproximadamente, 50 Hz. Os conjuntos de treino e teste da rede foram separados

através de inspeção visual. Duas modalidades de classificação foram consideradas: as 15 classes originais e outra, com apenas 4 classes, na qual os sinais foram identificados como pertencentes aos seguintes grupos: sintetizados, máquinas, biológicos e ruídos do oceano. O desempenho da rede FIRNN foi comparado com classificadores MLP e o discriminante generalizado de Fisher [49]. Para ambas modalidades, a rede FIRNN apresentou melhores resultados que os demais métodos, tendo se situado em 98,7% e 97,0% de generalização para os conjuntos de 4 e 15 classes, respectivamente.

Azimi-Sadjadi et al. (2000) [26] propuseram a classificação de ecos de sonar ativos através de um sistema extrator de características e de uma rede neural MLP, considerando duas classes distintas: minas e rochas. A extração de características foi baseada na combinação das técnicas de transformada *wavelet* [24] com a codificação preditiva linear (LPC - *Linear Predictive Coding*) [50]. O conjunto de dados foi coletado na *Coastal System Station*, na Flórida, EUA, num ambiente controlado (tanque de água), consistindo em 6 tipos diferentes de objetos, dos quais dois pertencem à classe mina e quatro, à classe rocha. A transformada *wavelet* foi utilizada para a decomposição do sinal em várias subbandas, das quais seis foram selecionadas. Para estas subbandas, um modelo de quarta ordem foi ajustado pelo método LPC, e parte dos coeficientes, selecionados pela função discriminante de Fisher [49], foram utilizados como características. A topologia ótima do classificador, a análise de robustez e a estimação do erro de classificação foram realizadas através de várias redes MLP, treinadas utilizando momento e taxa de aprendizado adaptativa [4]. A eficiência de generalização atingiu 92,7%. Foi constatado ainda que a combinação de múltiplos resultados de classificação produziu um ganho de eficiência, elevando-a para o valor de 97,7%, quando os resultados foram combinados linearmente, e para 99,0%, quando a combinação foi realizada por uma rede neural.

C. R Wan et al. (2000) [51] analisaram as propriedades do espectro de potência de sinais de sonar passivo, propondo um detetor ótimo pelo método clássico da máxima verossimilhança e por teste de hipóteses [52]. Foi suposto que o ruído aquático possuía uma distribuição gaussiana, uma vez que é normalmente constituído por sinais provenientes de múltiplas fontes, com fases e amplitudes aleatórias [53], o que resulta num espectro de potência com distribuição chi-quadrado [12]. O detetor foi avaliado com base em dados simulados (tons com adição de ruídos gaussianos)

e reais, o último obtido através de um ensaio de águas rasas, onde a fonte acústica produziu tons de freqüência constante.

Soares Filho (2001) [2], utilizando um banco de dados formado pela aquisição do ruído irradiado por quatro classes de navios distintas, na raia acústica da Marinha do Brasil, em Arraial do Cabo, desenvolveu um estudo detalhado sobre o pré-processamento necessário à classificação destes sinais. Esta análise foi baseada no conhecimento a priori das características deste ruído, sendo avaliado o impacto da resolução, da faixa de freqüência, da técnica de remoção de ruído de fundo e da normalização e média dos espectros na eficiência de classificação. Sistemas de classificação neurais foram propostos, sendo avaliado, inclusive, o impacto da redução de dimensionalidade dos sinais na capacidade de generalização do classificador. Para a classificação baseada em apenas uma janela espectral, uma eficiência de 92,2% de generalização foi obtida, valor que atingiu 99% quando a média de 30 espectros consecutivos foi aplicada à entrada do sistema. A classificação de espectros compactados explorou 3 técnicas de compactação: PCA, PCA não linear (NLPCA) [54] e componentes principais de discriminação (PCD) [55]. Para um total de 20 componentes, a análise PCA atingiu 87,0% de eficiência de generalização; enquanto que, para 9 componentes fornecidas pela NLPCA e 4 pela PCD, a eficiência obtida foi de 94%. Por fim, a inclusão de novas classes foi analisada, para a qual foi utilizada uma técnica de aprendizado não-supervisionada inspirada nas redes ART-2 [56].

Fernandez (2005) [57], utilizando um conjunto de dados formado por 8 classes de navios, de características similares ao utilizado por Soares Filho [2], propôs a classificação de espectros do sinal irradiado com base na técnica de curvas principais [58]. Foi proposta uma técnica de classificação baseada na distância de cada evento à curva representativa de sua classe, cuja extração era realizada pela técnica descrita em [59]. O critério de classificação adotado é o da mínima distância, ou seja, é atribuída ao evento a classe correspondente à curva mais próxima. Resultados expressivos foram obtidos, a despeito da simplicidade do critério de classificação, totalizando uma eficiência mínima e média de generalização de 93,5 % e 96,7 %, respectivamente.

2.1.2 Base de dados e pré-processamento

Uma opção interessante à constituição de uma base de dados que caracterize as diferentes classes é considerar um ambiente de medição controlado, onde um número finito de variáveis é monitorado. Procedimento comum consiste em realizar, com base nos navios disponíveis, medições em raia acústica. Neste procedimento, cada navio é submetido a um conjunto de condições de maquinário e operação conhecidas, e realiza um corrida, isto é, percorre a raia acústica, mantendo uma mesma configuração de operação, sendo seu ruído adquirido através de um hidrofone posicionado ao fundo da raia. Estas aquisições são portanto dependentes das condições físicas e oceanográficas da raia. A Figura 2.2 ilustra o processo descrito.

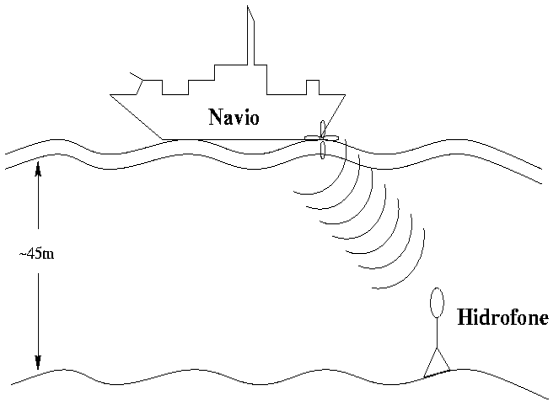


Figura 2.2: Ilustração do processo de aquisição de uma corrida numa raia.

A Marinha do Brasil, através de seu instituto de pesquisas (IPqM), disponibilizou um banco de dados para o desenvolvimento de trabalhos científicos, o qual foi utilizado neste trabalho. Este banco totaliza 263 corridas de 8 classes de navios distintas, adquiridas na sua raia acústica, situada em Arraial do Cabo. Cada corrida foi digitalizada utilizando uma freqüência de amostragem de 22,05 kHz e utilizou 16 *bits* para a quantização das amplitudes. Para cada classe, tem-se, ainda, um total de 2 a 5 navios, com, no mínimo, 4 corridas por navio.

Estudo detalhado sobre a estruturação do pré-processamento e uma análise do seu impacto na classificação de sinais de sonar passivo, para um conjunto de dados análogo ao considerado neste trabalho, foi realizado em [2]. Este estudo orientou

a escolha da cadeia de pré-processamento utilizada para o tratamento dos sinais envolvidos neste trabalho, a qual é apresentada na Figura 2.3.

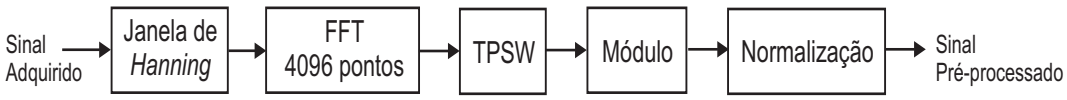


Figura 2.3: Diagrama em blocos do sistema de pré-processamento.

Através desta cadeia, o sinal adquirido é dividido em janelas de 4096 amostras, sem superposição, o que corresponde a uma janela de duração de, aproximadamente, 186 ms. Cada janela espectral foi submetida a um janelamento de Hanning [44], que visa minimizar os efeitos espectrais do recorte do sinal, e o módulo de sua transformada de Fourier foi determinado. Dos 4096 valores disponíveis, os 557 primeiros valores são considerados, os quais cobrem a faixa de 0 a $\approx$ 3 kHz, a uma resolução de $\approx$ 5,4 Hz por ponto espectral. Esta faixa de freqüência foi selecionada por concentrar as informações relevantes sobre as máquinas em operação no interior dos contatos [2]. Em seguida, os pontos espectrais são submetidos ao algoritmo TPSW (*Two-Pass Slit Window*) [60], o qual estima o ruído de fundo do ambiente de medição, já que este não contribui, de forma significativa, para a classificação dos contatos. Para esta estimativa, o algoritmo TPSW utiliza um janela de largura $2n + p$ pontos, possuindo uma fenda central de p pontos, conforme Figura 2.4 [1], a qual é utilizada para estimar o valor médio dos pontos comprendidos pela janela.

A operação do algoritmo TPSW sobre o espectro se processa em duas etapas: numa primeira, é realizada a convolução da janela com cada ponto do espectro, provendo uma média local inicial. Esta média é multiplicada por um fator α, e comparada com um limiar. Caso o limiar seja excedido, os pontos correspondentes são substituídos por esta média local. De posse deste espectro modificado, uma segunda convolução é realizada, resultando na estimativa final da média local, a qual é utilizada para a divisão, ponto-a-ponto, dos pontos espectrais provenientes do bloco módulo. Os parâmetros utilizados foram $n = 50$, $p = 4$ e $\alpha = 2.0$, e

[1]Nesta figura: p=5 e n=6.

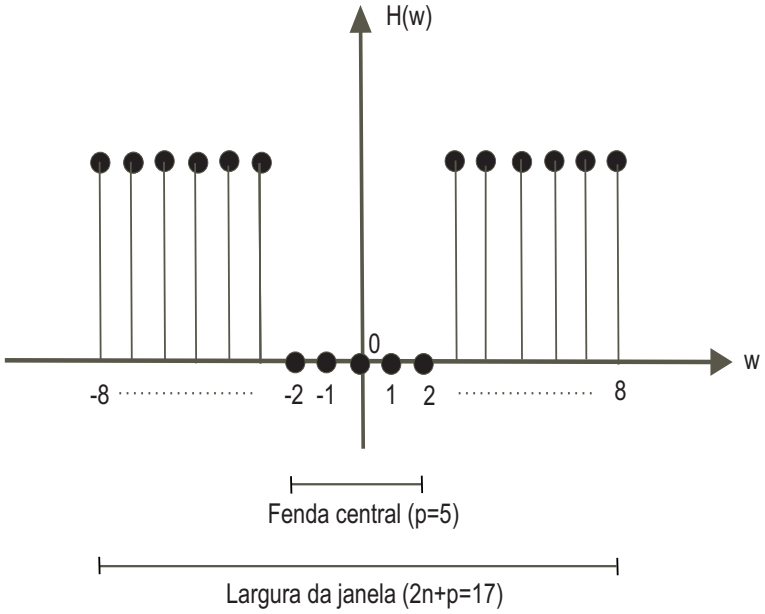


Figura 2.4: Janela utilizada pelo algoritmo TPSW para a estimativa do ruído de fundo do ambiente de medição.

sua escolha foi baseada em resultados apresentados em [2]. Por fim, é realizada um normalização do espectro resultante para que apresente energia unitária.

Na Tabela 2.1 é apresentada a composição de cada classe em termos de navios e corridas. Para cada classe, são apresentados os navios constituintes, juntamente com o ano de aquisição, o total de corridas disponíveis (N.C.) e o número de janelas espectrais produzidas pela cadeia de pré-processamento (N.E.). Dispõe-se, portanto, de, no mínimo 19 (classe G), e no máximo 66 (classe E) corridas para a caracterização de cada classe. Em termos de janelas espectrais, tem-se, no mínimo 2143, e no máximo, 7075 janelas, totalizando, para todas as classes, 29277 janelas. Pode-se verificar que o conjunto utilizado possui um número significativo de eventos de dimensão elevada, o que restringe as técnicas aplicáveis à solução do problema e dificulta o projeto do classificador.

2.1.3 Considerações sobre o sistema de classificação

Para a produção e a avaliação do sistema de classificação discutido neste trabalho, dois enfoques foram considerados: um primeiro, onde o sistema é avaliado considerando o mesmo conjunto de condições operativas utilizado em seu desenvol-

Tabela 2.1: Distribuição das corridas por classes e navios (Veja o texto).

Classe A				**Classe B**			
Navio	**Ano**	**N.C.**	**N.E.**	**Navio**	**Ano**	**N.C.**	**N.E.**
A1	97	11	1191	B1	90	8	914
A2	95	4	504	B2	88	4	700
A4	97	7	737	B3	92	9	1224
				B5	89	5	594
Total corridas: 22		**Total espectros:** 2432		**Total corridas:** 26		**Total espectros:** 3432	
Classe C				**Classe D**			
Navio	**Ano**	**N.C.**	**N.E.**	**Navio**	**Ano**	**N.C.**	**N.E.**
C1	90	5	512	D1	95	13	1401
C2	96	10	1081	D3	98	4	406
C3	91	6	799	D4	90	12	1265
C4	89	6	705				
C5	97	16	1700				
Total corridas: 43		**Total espectros:** 4797		**Total corridas:** 29		**Total espectros:** 3072	
Classe E				**Classe F**			
Navio	**Ano**	**N.C.**	**N.E.**	**Navio**	**Ano**	**N.C.**	**N.E.**
E1	92	19	2202	F1	97	6	636
E1	97	2	260	F2	97	8	857
E2	95	7	717	F3	97	6	628
E3	91	6	671	F4	97	7	813
E3	93	25	2834				
E4	97	7	391				
Total corridas: 66		**Total espectros:** 7075		**Total corridas:** 27		**Total espectros:** 2934	
Classe G				**Classe H**			
Navio	**Ano**	**N.C.**	**N.E.**	**Navio**	**Ano**	**N.C.**	**N.E.**
G1	90	8	867	H1	89	16	1680
G2	90	11	1276	H2	88	9	975
				H5	85	6	737
Total corridas: 19		**Total espectros:** 2143		**Total corridas:** 31		**Total espectros:** 3392	

vimento; e um segundo, onde a avaliação pode considerar condições operativas não contempladas no projeto do classificador. Para o primeiro caso, como cada corrida responde por uma condição operativa específica, as etapas de projeto e avaliação consideram todas as corridas disponíveis, utilizando, no entanto, diferentes conjuntos de janelas espectrais. Para o segundo, cada etapa considera um conjunto diferente de corridas. Estas duas modalidades serão aqui referidas como projeto e avaliação baseados em espectros e corridas, respectivamente.

Para cada modalidade, um diferente teste da capacidade de generalização e robustez do classificador é realizado. Na modalidade por corridas, tem-se um teste mais severo da capacidade de generalização do classificador, visto que um contato de classe conhecida, porém com condições operativas possivemente não exploradas no processo de treinamento, é submetido ao classificador. A avaliação baseada em corridas é, portanto, mais realística com respeito à operação real do sistema, tendo em vista a multiplicidade de cenários possíveis no contexto do sonar. Por outro lado, para as classes mais críticas, a exclusão de condições operativas no projeto do classificador pode comprometer, em demasia, a caracterização estatística das classes. Assim, uma melhor caracterização estatística do que é conhecido é obtida pela modalidade por espectros.

Dado que os sinais foram obtidos no interior de uma raia acústica, com o navio razoavelmente próximo do sensor, eles não correspondem exatamente aos que seriam encontrados numa operação real, possuindo relação sinal-ruído superior ao esperado. Num cenário real de operação, os sinais captados pelo sistema contariam, ainda, com flutuações estatísticas relacionadas ao efeito de diferentes condições oceanográficas, do ruído produzido pelo próprio submarino, e com a possibilidade de mais de um contato simultâneo. O estudo realizado neste trabalho pode, no entanto, orientar novos desenvolvimentos, onde medições controladas, realizadas como o submarino em operação, estejam disponíveis.

Parte II

Pré-processamento e compactação

Capítulo 3

Métodos adaptativos para a extração de componentes principais

Conforme seção 2.1, para o sistema de classificação automática de contatos, um módulo bastante relevante ao desempenho do sistema é o extrator de características. Em geral, sistemas extratores de características realizam uma redução da dimensão dos vetores de dados, retendo ou enfatizando as variáveis que estão correlacionadas à solução do problema. De posse de um menor número de variáveis, reduz-se a complexidade do espaço de dados, o que resulta em classificadores de maior eficiência. Outra conquista é a simplificação do processo de aprendizado, dado que o número de parâmetros a serem estimados no treinamento é reduzido, o que é de extrema valia quando os dados possuem restrições quanto sua caracterização estatística. A redução da complexidade do classificador resulta, também, num menor custo computacional, o que é relevante, ou mesmo fundamental, para sua implementação em tempo real.

Um sistema de classificação que incorpore novos cenários exige uma capacidade de adaptação do extrator, o que restringe as técnicas aplicáveis. Uma técnica de extração que conjuga capacidade de adaptação, eficácia e simplicidade é a análise de componentes principais (PCA), descrita em maiores detalhes no Apêndice B. As componentes principais fornecem direções privilegiadas para a representação dos dados, provendo uma eficiente redução da dimensão dos dados, e realizam uma

compactação baseada em energia [4].

Há um número expressivo de algoritmos na literatura para a extração de componentes principais. Em linhas gerais, tem-se as técnicas clássicas, propostas pela comunidade da álgebra linear; e as adaptativas, derivadas pelas comunidades de redes neurais e de processamento de sinais. As técnicas clássicas, a despeito de acuradas, são, em geral, contra-indicadas, ou mesmo inviáveis, em aplicações de alta-dimensionalidade e ambientes não-estacionários, dada as exigências de armazenamento e processamento impostas pelo cálculo da matriz de correlação dos dados. Os algoritmos adaptativos, por outro lado, permitem a extração de um número arbitrário de componentes, segundo um nível de acuidade definido para a aplicação, não exigindo, para a maior parte dos algoritmos, o cálculo da matriz de correlação dos dados.

Na literatura verifica-se um número significativo de propostas de algoritmos adaptativos para PCA, em especial utilizando redes neurais artificiais lineares. Para um problema particular, surgem várias perguntas, entre elas: quais métodos estão disponíveis? qual deles utilizar numa aplicação específica? quais são as diferenças entre dois dados algoritmos da literatura?

Pequeno é o número de trabalhos onde uma revisão bibliográfica sobre métodos adaptativos para a análise PCA é realizada. Em [4], apenas três métodos são citados: Oja (1982) [61], Sanger (1989) [38] e APEX (1994) [62]. Uma revisão expressiva de técnicas clássicas para a extração de um pequeno número de componentes é Comon e Golub (1990) [63]. Em Kung e Diamantaras (1996) [64], tem-se uma revisão abrangente, que carece, no entanto, dos algoritmos mais recentes, de maior velocidade de convergência. Outra revisão é encontrada em Baldi e Hornik(1995) [65], bastante detalhada, que enfoca algoritmos *on-line* baseados em princípios hebianos. Em Hyvärinen et al.(2001) [66] são destacadas técnicas baseadas no princípio hebbiano. Um enfoque unificado para a extração de componentes minoritárias e majoritárias, de processos aleatórios reais e complexos, é apresentada em [67].

Ainda que as referências [64, 65, 67] possam ser utilizadas para uma primeira seleção de algoritmos, as referências [64, 65] carecem de técnicas mais recentes e de algoritmos de extração *off-line*, e a referência [67] não discute as especificidades e relações existentes entre os diferentes algoritmos apresentados. Desta forma, uma

escolha mais criteriosa de um algoritmo, para uma aplicação em particular, exige uma pesquisa entre os vários periódicos disponíveis, cada qual com uma notação algébrica e forma de apresentação próprias. Também são raros os trabalhos onde o desempenho de um grupo de algoritmos é avaliado, em especial, com base em dados reais. Mesmo nos artigos onde uma análise comparativa é realizada [68], não é comum verificar uma análise mais detalhada das topologias utilizadas que justifique as diferenças de desempenho encontradas nos testes realizados.

Grande parte das técnicas apresentadas na literatura é baseada em gradiente estocástico, técnica de otimização simples, freqüentemente utilizada para a produção de algoritmos *on-line* [65], porém de baixa velocidade de convergência [4]. Há uma carência de trabalhos que proponham métodos eficientes para a extração *off-line* de um número arbitrário de componentes, em especial quando são consideradas aplicações de alta-dimensionalidade e grande número de eventos. Ainda que a referência [63] reúna técnicas para a extração de um número pequeno de componentes, os algoritimos apresentados exigem o cálculo da matriz de correlação dos dados, o que é contra-indicado, ou mesmo inviável, para um número significativo de aplicações de alta-dimensionalidade.

A pesquisa de algoritmos adaptativos de extração PCA aplicáveis ao problema de sonar passivo motivou a elaboração do presente capítulo, que possui duas contribuições principais: a primeira é prover uma revisão dos principais algoritmos disponíveis na literatura. Esta revisão buscou apresentar uma visão intuitiva dos algoritmos, agrupado-os em famílias, de acordo com as estratégias exploradas por cada método. Buscou-se, portanto, fornecer uma orientação com respeito a escolha de uma técnica mais adequada à aplicação em um problema particular, em detrimento de uma análise teórica mais detalhada sobre as propriedades de convergência de cada algoritmo. A segunda contribuição é propor um conjunto de métodos de extração de componentes principais, cada qual mais adequado a um domínio de aplicação (*on-line* x *off-line*), baseados no treinamento de uma arquitetura neural aqui proposta, batizada como hierarquia de células auto-associativas. Relações entre os métodos propostos e algoritmos da literatura são também derivadas.

A revisão da bibliográfica dos algoritmos buscou contemplar trabalhos recentes, em especial, técnicas de baixo custo computacional por iteração. A apresentação

é baseada numa divisão didática do assunto em dois tópicos: funções objetivo e topologias neurais de extração. Através desta divisão, é possível prover uma melhor compreensão sobre o desempenho e os domínios de aplicação de cada técnica. São destacados os algoritmos de maior velocidade de convergência, visando atender as exigências de aplicações de alta-dimensionalidade e grande número de eventos, tais como o caso do sonar passivo. Outra contribuição é que os diferentes algoritmos são apresentados segundo uma notação unificada, o que facilita a compreensão, implementação e comparação das diferentes técnicas.

Com respeito aos algoritmos propostos, inicialmente é apresentada a hierarquia de células auto-associativas, que é seguida pela apresentação dos métodos de treinamento desta estrutura. Após, é discutida a extração de componentes principais do conjunto de dados de sonar passivo, sendo realizada uma comparação de acuidade e custo computacional entre os métodos propostos e métodos consagrados da literatura, tais como o GHA [38] e o PAST [69].

3.1 Algoritmos adaptativos para extração de componentes

Uma das principais características de um algoritmo adaptativo é basear-se na seguinte equação recursiva:

$$\mathbf{w}(k+1) = F(\mathbf{w}(k), \mathbf{x}_s), \tag{3.1}$$

onde o valor do parâmetro em adaptação é representado por $\mathbf{w}$, a iteração corrente corresponde à k e o valor futuro do parâmetro (iteração $k+1$) é determinado por uma função arbitrária do parâmetro atual $\mathbf{w}(k)$ e de um subconjunto dos dados $\mathbf{x}_s$, de tamanho arbitrário. Para sistemas neurais, a função F é conhecida como equação de treinamento, $\mathbf{w}$ representa os pesos e limiares e $\mathbf{x}_s$ é o subconjunto dos dados considerados na k-ésima iteração de treinamento da rede.

Os algoritmos adaptativos podem ainda ser classificados em *on-line* e *off-line*. Para os algoritmos *on-line*, a cada evento disponível, é produzida uma modificação nos parâmetros [65]. Por outro lado, os algoritmos *off-line* exigem o armazenamento de vários eventos, tipicamente todo conjunto de dados, cuja composição define a

atualização. Quanto a cardinalidade de $\mathbf{x}_s$, em geral ela é 1 e n para os algoritmos *on-line* e *off-line*, respectivamente.

Sistemas para os quais os dados não estão disponíveis a priori exigem algoritmos *on-line*. Para aplicações onde os dados foram previamente coletados utilizam-se, comumente, algoritmos *off-line*. Para algumas aplicações é possível adaptar técnicas *off-line* de forma a operarem numa forma "*on-line* aproximada", a qual é referida como *real-time* (tempo real).

Algoritmos adaptativos para a extração de componentes principais são, normalmente, derivados pela otimização de funções objetivo, cujos extremos estão relacionados às componentes. Como conseqüência, podem ser produzidos algoritmos que extraem as componentes propriamente ditas ou um subespaço [70], dentre os vários possíveis, por elas determinado. No primeiro caso, tem-se uma extração das componentes principais, o que é conhecido como PCA (*Principal Component Analysis*); no segundo, uma extração das componentes do subespaço, referida como PSA (*Principal Subspace Analysis*).

A otimização de funções objetivo para a extração de componentes principais é realizada, freqüentemente, por técnicas baseadas em gradiente descendente ou ascendente [71]. Grande é o número de trabalhos que utiliza o gradiente estocástico [4], originando algoritmos *on-line*. A utilização de técnicas mais sofisticadas de otimização é mais freqüente em algoritmos *off-line*. Cabe observar que o binômio função objetivo e algoritmo de otimização possui forte relação com a velocidade de convergência do algoritmo. Quanto "menos plana"for a superfície associada à função objetivo e mais sofisticado for o algoritmo de otimização, menor será o número de iterações até a convergência. Para sistemas *on-line*, um dos principais impactos da velocidade de convergência é a capacidade do sistema em se ajustar a mudanças do ambiente de operação. Para sistemas *off-line*, o principal efeito a ser considerado é o custo computacional, que normalmente se traduz em tempo de extração.

Número expressivo de trabalhos constatou que a extração de componentes principais pode ser realizada pelo treinamento de redes neurais quando utilizadas funções objetivo apropriadas. Os algoritmos neurais de extração PCA podem ser agrupados em duas famílias, de acordo com a estratégia utilizada: as técnicas baseadas em redes auto-associativas e na maximização da variância, ambas detalhadas

posteriormente.

3.2 Funções de custo para a extração de componentes

Conforme discussão inicial, dadas as características da função objetivo e da técnica de otimização utilizada para sua obtenção, os algoritmos apresentam diferentes velocidades de convergência, níveis de complexidade de implementação e exigências quanto a capacidade de armazenamento.

Na seção 3.1 é mostrado que a expansão de um vetor aleatório $\mathbf{x}$ em p componentes principais pode ser escrita como:

$$\tilde{\mathbf{x}}_p = \mathbf{Q}_p \mathbf{Q}_p^T \mathbf{x}, \tag{3.2}$$

onde $\mathbf{Q}_p$ é uma matriz com colunas correspondentes às p componentes principais do vetor $\mathbf{x}$. O erro médio cometido nesta expansão é mínimo, sendo dado por:

$$EMQ = E[(\mathbf{x} - \tilde{\mathbf{x}_\mathbf{p}})^2] = E[(\mathbf{x} - \mathbf{Q}_p \mathbf{Q}_p^T \mathbf{x})^2] \tag{3.3}$$

Deste modo, uma possível função objetivo, freqüentemente utilizada, é o erro médio de reconstrução, o qual é dado por:

$$EQ = E[(\mathbf{x} - \mathbf{W}_p \mathbf{W}_p^T \mathbf{x})^2], \tag{3.4}$$

onde $\mathbf{W}_p$ é uma matriz n x p correspondente aos parâmetros a serem obtidos pela otimização da função EQ.

O valor ótimo de $\mathbf{W}_p^*$ na Equação 3.4 é dado por:

$$\mathbf{W}_p^* = \mathbf{Q}_p \mathbf{T}, \tag{3.5}$$

onde $\mathbf{T}$ é uma matriz ortonormal ($\mathbf{T}\mathbf{T}^T = \mathbf{I}$). Vale observar que a Equação 3.4 for minimizada sem a utilização de qualquer estratégia complementar, não é garantido que $\mathbf{T} = \mathbf{I}$, sendo produzido um algoritmo de extração das componentes do subespaço principal.

Outra função comum na literatura é:

$$F = tr[\mathbf{W}_p^T \mathbf{R}_\mathbf{x} \mathbf{W}_p], \tag{3.6}$$

onde $\mathbf{R_x} = E[\mathbf{x}\mathbf{x}^T]$. Esta equação pode ainda ser escrita como:

$$F = tr[\mathbf{W}_p^T \mathbf{R_x} \mathbf{W}_p] = E[tr\{\mathbf{c}\mathbf{c}^T\}] = \sum_{i=1}^{p} E[c_i^2], \tag{3.7}$$

para $\mathbf{c} = [c_1 \dots c_p]$ dado por: $c_i = \mathbf{w}_i^T \mathbf{x}$, e $\mathbf{w}_i$ correspondente à i-ésima coluna de $\mathbf{W}_p$, ou seja, realiza-se a otimização da energia das projeções, o que equivale, para processos de média nula, a otimização da variância.

Retomando a Equação 3.4, tem-se [1]:

$$\begin{aligned} E &= E[(\mathbf{x} - \mathbf{W}_p\mathbf{W}_p^T\mathbf{x})^2] = tr\{E[(\mathbf{x} - \mathbf{W}_p\mathbf{W}_p^T\mathbf{x})(\mathbf{x} - \mathbf{W}_p\mathbf{W}_p^T\mathbf{x})^T]\} \\ E &= tr\{E[\mathbf{x}\mathbf{x}^T + 2\mathbf{W}_p\mathbf{W}_p^T\mathbf{x}\mathbf{x}^T(\mathbf{W}_p\mathbf{W}_p^T - \mathbf{I}) - \mathbf{W}_p\mathbf{W}_p^T\mathbf{x}\mathbf{x}^T\mathbf{W}_p\mathbf{W}_p^T]\} \\ E &= tr\{E[\mathbf{x}\mathbf{x}^T]\} - tr\{\mathbf{W}_p^T E[\mathbf{x}\mathbf{x}^T]\mathbf{W}_p\} \\ E &= tr\{\mathbf{R_x}\} - tr\{\mathbf{W}_p^T\mathbf{R_x}\mathbf{W}_p\}, \end{aligned} \tag{3.8}$$

ou seja, caso as colunas de $\mathbf{W}_p$ sejam ortonormais, as Equações 3.4 e 3.6 possuem pontos ótimos idênticos. Deste modo, tanto a minimização da Equação 3.4 quanto a maximização da Equação 3.6 produzem as componentes do subespaço. Cabe destacar que na maximização da Equação 3.6, para evitar um crescimento indefinido das colunas de $\mathbf{W}_p$, é necessário impor restrições quanto a norma das estimativas das componentes [72].

3.3 Funções de custo alternativas para a extração de componentes principais

Visando otimizar o processo de extração (convergência mais rápida) ou realizar uma extração das componentes principais, algumas funções objetivo alternativas foram propostas, entre elas:

3.3.1 Brockett

Brockett (1989/1991) [73, 74] propôs a função:

$$J = tr[\mathbf{W}_p\mathbf{R_x}\mathbf{W}_p^T\mathbf{A}_p], \tag{3.9}$$

[1] $tr\{\mathbf{A}\}$ designa o operador traço da matriz $\mathbf{A}$ [70]. Nesta dedução foi explorada a propriedade: $tr\{\mathbf{AB}\} = tr\{\mathbf{BA}\}$ e o fato das colunas de $\mathbf{W}$ serem supostamente ortogonais.

onde $\mathbf{W}_p$ é a matriz de parâmetros, de dimensões p x n, onde p indica o número de componentes a extrair. Esta equação é análoga à Equação 3.6, exceto pela matriz diagonal $\mathbf{A}_p$, de dimensões p x p, cujos valores devem ser constantes decrescentes. O autor mostra que a otimização desta função com respeito à matriz $\mathbf{W}_p$ permite a obtenção das componentes principais nas suas linhas. Cabe observar que a convergência não se faz de forma ordenada, logo duas otimizações que partam de valores de pesos iniciais distintos podem produzir diferentes distribuições das componentes nas linhas da matriz $\mathbf{W}_p$.

De forma análoga à Equação 3.6, a Equação 3.9 pode ser escrita na forma:

$$J = \sum_{i=1}^{p} \theta_i E[c_i^2], \tag{3.10}$$

onde θ_i é um conjunto de constantes decrescentes correspondente aos elementos das diagonal da matriz $\mathbf{A}_p$ e $c_i = \mathbf{w}_i^T \mathbf{x}$. Tem-se, pois, uma otimização ponderada da variância das projeções, onde cada componente contribui de forma diferenciada, determinada pelo valor de θ_i.

3.3.2 Lei Xu

Lei Xu (1993) [75] propôs para o treinamento de uma arquitetura neural própria, a ser descrita posteriormente (vide seção 3.4.1), a função:

$$J = \frac{1}{2}.E[||\mathbf{x} - \mathbf{W}_p \mathbf{A}_p \mathbf{W}_p^T \mathbf{x}||^2], \tag{3.11}$$

para $\mathbf{W}_p$ com dimensão (n x p), onde p é o número de componentes a extrair, e a matriz $\mathbf{A}_p$ é idêntica àquela considerada pela Equação 3.9. Outra contribuição do trabalho é mostrar que as Equações 3.11 e 3.9 possuem os mesmos mínimos globais, de forma similar à relação estabelecida entre as Equações 3.4 e 3.6, e correspondem a superfícies sem mínimos locais, apenas pontos de sela.

3.3.3 Coeficiente de Rayleigh

Número razoável de técnicas é baseada na otimização do coeficiente de Rayleigh, que é dado por [76]:

$$R = \frac{\mathbf{w}(k)^T \mathbf{A} \mathbf{w}(k)}{\mathbf{w}(k)^T \mathbf{w}(k)}, \tag{3.12}$$

onde $\mathbf{A}$ é uma matriz arbitrária. O valor máximo de R ocorre quando $\mathbf{w} = \pm\alpha\mathbf{e}_1$, onde α é uma constante arbitrária, e $\mathbf{e}_1$ corresponde ao autovetor dominante de $\mathbf{A}$. Assim, a maximização de R considerando $\mathbf{A} = \mathbf{R}_\mathbf{x}$ pode ser utilizada para a extração de componentes principais. Caso seja imposto que $|\mathbf{w} = 1|$, através da normalização do vetor $\mathbf{w}$ a cada iteração ou por uma restrição de otimização, a Equação 3.12 pode ser escrita como:

$$R_a = \mathbf{w}(k)^T \mathbf{A}\mathbf{w}(k), \tag{3.13}$$

a qual é similar à Equação 3.6 quando $p = 1$.

Para a extração de p componentes, uma estratégia é realizar a minimização de p funções objetivo que contenham a Equação 3.13. Nesta minimização, faz-se necessário impor duas restrições: quanto à norma e quanto à ortogonalidade dos pesos. Através da primeira, evita-se um crescimento indefinido da norma de $\mathbf{w}$. Pela segunda, evita-se que todas as minimizações obtenham a mesma componente principal (a primeira).

Utilizando a técnica de multiplicadores de Lagrange [77], a função objetivo cuja minimização realiza a extração da j-ésima componente prinicipal pode ser escrita como [78]:

$$\begin{aligned} J_j = -\mathbf{w}_j^T(k)\mathbf{R}_\mathbf{x}\mathbf{w}_j(k) + \alpha_j(k)[1 - \mathbf{w}_j^T(k)\mathbf{w}_j(k)] \\ + k\sum_{i=1}^{t} \beta_{ji}(k)\mathbf{w}_j^T(k)\mathbf{w}_i(k), \qquad 1 \leq j \leq p, \end{aligned} \tag{3.14}$$

onde $\alpha_j(k)$ e $\beta_{ji}(k)$ são multiplicadores de Lagrange a serem determinados, t é um inteiro, e $\mathbf{w}_j$ corresponde à componente a ser extraída. Nesta equação, três termos compõem a função a ser minimizada: o primeiro, relacionado à Equação 3.13; o segundo, que será nulo quando a norma de $\mathbf{w}_j$ for unitária; e o último, nulo, se o peso $\mathbf{w}_j$ for ortogonal aos pesos de índices de 1 a t.

Calculando o gradiente de J_j em relação a $\mathbf{w}_j(k)$, tem-se:

$$\nabla_{\mathbf{w}_j(k)} J_j = -2\mathbf{R}_\mathbf{x}\mathbf{w}_j(k) - 2\alpha_j(k)\mathbf{w}_j(k) + k\sum_{i=1}^{t} \beta_{ji}(k)\mathbf{w}_i(k) \tag{3.15}$$

Se multiplicarmos o gradiente de J_j à esquerda por $\mathbf{w}_j(k)^T$ e igualarmos o resultado a zero, considerando as restrições $\mathbf{w}_j(k)^T\mathbf{w}_i(k) = 0$, para $i \neq j$, e

$|\mathbf{w}_j(k)| = 1$, resulta que o multiplicador $\alpha_j(k)$ é dado por:

$$\alpha_j(k) = \mathbf{w}_j^T(k)\mathbf{R}_\mathbf{x}\mathbf{w}_j(k) \tag{3.16}$$

Similarmente, multiplicando o gradiente de J_j à esquerda por $\mathbf{w}_i(k)^T$ e igualando o resultado a zero, para as mesmas restrições, resulta:

$$\beta_{ji}(k) = \frac{2}{k}\mathbf{w}_i(k)^T\mathbf{R}_\mathbf{x}\mathbf{w}_j(k), \qquad i = 1, \ldots, t \tag{3.17}$$

Substituindo as Equações 3.16 e 3.17 na Equação 3.14, resulta:

$$\begin{aligned} J_j = &-\mathbf{w}_j^T(k)\mathbf{R}_\mathbf{x}\mathbf{w}_j(k)\mathbf{w}_j^T(k)\mathbf{w}_j(k) \\ &+ 2\sum_{i=1}^{t}\mathbf{w}_i^T(k)\mathbf{R}_\mathbf{x}\mathbf{w}_j(k)\mathbf{w}_j^T(k)\mathbf{w}_i(k), \qquad 1 \leq j \leq p \end{aligned} \tag{3.18}$$

Esta estratégia foi explorada pelos trabalhos de Karhunen e Joutsensalo (1995) [79] e Chatterjee, Kang e Roychowdhury (2000) [78]. É mostrado em [79] que, caso t seja escolhido como $t = j - 1$, a minimização da Equação 3.18 realiza a extração das componentes principais; caso $t = n$, é realizada a extração das componentes do subespaço.

3.3.4 NIC - Novel Information Criteria

O algoritmo NIC, proposto por Mao e Hua (1998) [80], utiliza a seguinte função objetivo:

$$J_{NIC}(\mathbf{W}) = \frac{1}{2}tr[log(\mathbf{W}^T\mathbf{R}_\mathbf{x}\mathbf{W}) - tr(\mathbf{W}^T\mathbf{W})], \tag{3.19}$$

a qual é constituída por dois termos: o primeiro, análogo a Equação 3.6; e o segundo, responsável pela ortonormalização dos pesos da rede. Segundo os autores, o diferencial da técnica é o logaritmo, o qual resulta numa mudança da conformação da superfície que está sendo otimizada, resultando num significativo aumento da velocidade de convergência do algoritmo. A otimização da função objetivo NIC produz uma extração das componentes do subespaço.

3.3.5 WNIC - Weighted Novel Information Criteria

De autoria de Outang e Bao (2002) [81], consiste numa extensão do algoritmo NIC para a extração das componentes principais. A função objetivo proposta é dada

por:

$$J_{WNIC}(\mathbf{W}) = \frac{1}{2}\{tr[log(\mathbf{W}^T\mathbf{R_x}\mathbf{WA}) - tr(\mathbf{W}^T\mathbf{W})]\}, \tag{3.20}$$

a qual é análoga à utilizada pelo método NIC, exceto pela assimetria inserida pela matriz diagonal $\mathbf{A}$, cujos valores devem ser constantes decrescentes. É fácil perceber que a função proposta explora estratégia similar a utilizada no trabalho de Brockett [73] (vide seção 3.3.1).

3.3.6 Funções RLS

Esta linha de trabalhos é baseada em funções objetivo recursivas, análogas à função *Recursive Least Square* (RLS) utilizada na área de filtragem adaptativa [76].

No algoritmo de Bannour e Azimi-Sadjadi (1995) [82], para a extração de p componentes principais, é realizada a otimização de p funções objetivo dadas por:

$$J_j(n) = \sum_{k=1}^{n} ||\mathbf{d}_j(k) - h_j(k)\mathbf{w}_j(k-1)||^2, \qquad j = 1 \ldots p, \tag{3.21}$$

para:

$$h_j(k) = \mathbf{w}_j^T(k-1)\mathbf{x}(k), \tag{3.22}$$

e $\mathbf{d}_j(k)$ dado por:

$$\mathbf{d}_j(k) \quad = \mathbf{x}(k) - \textstyle\sum_{i=1}^{j-1} h_i(k)\mathbf{w}_i(k-1) \tag{3.23}$$

Pode-se notar que como $\mathbf{d}_j$ é dependente de $\mathbf{w}_i$, é imposta um relação entre as p funções objetivo a serem otimizadas.

Desenvolvendo-se a Equação 3.21, obtém-se uma equação de 4^o grau em $\mathbf{w}_j$, para a qual não é possível a determinação analítica do valor de $\mathbf{w}_j(k)$ que produz seu mínimo. Se, no entanto, esta equação for aproximada para:

$$J_j(n) \approx \sum_{k=1}^{n} ||\mathbf{d}_j(k) - h_j(k)\mathbf{w}_j(n-1)||^2, \qquad j = 1 \ldots p, \tag{3.24}$$

resulta numa equação de 2^o grau em $\mathbf{w}_j(n-1)$, cujo mínimo pode ser determinado analiticamente. Como o valor ótimo dos pesos pode ser determinado a cada iteração, produz-se um algoritmo de velocidade de convergência significativamente maior que

o obtido pela minização da Equação 3.21 por outra técnica de otimização (gradiente descendente, por exemplo), o que justifica a aproximação realizada.

Outra proposta deve-se a Yang (1995) [69], que propõe o algoritmo PAST (*Projection Approximation Subspace Tracking*), o qual realiza a extração de componentes do subespaço. O algoritmo parte da minimização da seguinte função objetivo:

$$J_{\mathbf{W}(t)} = \sum_{i=1}^{t} \beta^{t-i} ||\mathbf{x}(i) - \mathbf{W}(t)\mathbf{W}^T(t)\mathbf{x}(i)||^2, \tag{3.25}$$

onde β é uma constante arbitrária, que deve estar compreendida na faixa de 0 a 1, e $\mathbf{W}$ é uma matriz n x p, onde n é a dimensão dos dados de entrada e p corresponde ao número de componentes que se deseja extrair. Como a Equação 3.25 considera os dados da primeira até a iteração atual (t), o papel da constante β é introduzir um "esquecimento", isto é, a medida que cresce o número de iterações, reduz-se a contribuição das amostras mais antigas na composição de $J_{\mathbf{W}(t)}$.

De forma análoga ao algoritmo de Bannour [82], para permitir a determinação do valor ótimo de $\mathbf{W}$ a cada passo, é proposta a aproximação:

$$\mathbf{W}^T(t)\mathbf{x}(i) \approx \mathbf{W}^T(i-1)\mathbf{x}(i), \qquad 1 \leq i \leq t, \tag{3.26}$$

ou seja, considera-se que a variação da projeção dos dados no intervalo de iterações de $1 < i < t$ não seja significativa, o que resulta:

$$J_{\mathbf{W}(t)} \approx \sum_{i=1}^{t} \beta^{t-i} ||\mathbf{x}(i) - \mathbf{W}(t)\mathbf{y}(i)||^2, \tag{3.27}$$

onde:

$$\mathbf{y}(i) = \mathbf{W}(i-1)^T\mathbf{x}(i) \tag{3.28}$$

Em razão da aproximação de projeção realizada, verifica-se que as colunas de $\mathbf{W}(t)$ não são ortogonais após a convergência. Abed-Meraim, Chkeif e Hua (2000) [83] propuseram um método, de baixo custo computacional, para realizar, a cada passo de treinamento, a ortonormalização de $\mathbf{W}$.

Na tabela a seguir, resumimos as funções objetivo discutidas, apresentando o tipo de otimização (Otim.) envolvido (Mín - minimização e Máx - maximização), as principais referências onde a função foi proposta ou explorada e a modalidade de extração (Mod.) que é realizada (PSA ou PCA).

Tabela 3.1: Principais funções objetivo para extração PCA/PSA (Veja o texto).

Função objetivo	**Otim.**	**Referências**	**Mod.**
$J = tr[\mathbf{W}_p \mathbf{R}_\mathbf{x} \mathbf{W}_p^T \mathbf{A}_p]$	Máx.	Brockett (1989/1991) [73, 74]	PCA
$J = \frac{1}{2} E[\|\|\mathbf{x} - \mathbf{W}_p \mathbf{A}_p \mathbf{W}_p^T \mathbf{x}\|\|^2]$	Mín.	Lei Xu (1993) [75]	PCA
$J_j(n) = \sum_{k=1}^{n} \|\|\mathbf{d}_j(k) - h_j(k)\mathbf{w}_j(n-1)\|\|^2$	Mín.	Bannour (1995) [82]	PCA
$J_{\mathbf{W}(t)} = \sum_{i=1}^{t} \beta^{t-i} \|\|\mathbf{x}(i) - \mathbf{W}(t)\mathbf{y}(i)\|\|^2$	Mín.	Yang (1995) [69]	PSA
Equação 3.18	Máx.	Karhunen (1995) [79] Chatterjee (2000) [78]	PSA / PCA
$J(\mathbf{W}) = \frac{1}{2} tr[log(\mathbf{W}^T \mathbf{R}_\mathbf{x} \mathbf{W}) - tr(\mathbf{W}^T \mathbf{W})]$	Máx.	Mao e Hua (1998) [80]	PSA
$J(\mathbf{W}) = \frac{1}{2} \{tr[log(\mathbf{W}^T \mathbf{R}_\mathbf{x} \mathbf{W}\mathbf{A}) - tr(\mathbf{W}^T \mathbf{W})]\}$	Máx.	Outang e Bao (2002) [81]	PCA

3.4 Arquiteturas neurais para a extração de componentes

Número significativo de algoritmos de extração de componentes é baseado no treinamento de redes neurais com arquiteturas e funções objetivo específicas. Em linhas gerais, os algoritmos podem ser agrupados nas seguintes linhas de raciocínio:

3.4.1 Redes auto-associativas

Esta família compreende redes de duas camadas de neurônios lineares, treinadas para realizar um mapeamento identidade [64], ou seja, reproduzir na saída o vetor de entrada. Para este fim, uma arquitetura comumente utilizada é a MLP [4]. Neste caso, para dados de dimensão n, utiliza-se uma rede de dimensões n x p x n, onde $p \leq n$, ou seja, impõe-se um gargalo na camada intermediária [64]. Quanto ao treinamento, os vetores-alvo são considerados idênticos aos dados de entrada, o que é conhecido como treinamento não-supervisionado auto-associativo [65]. Na Figura 3.1 é ilustrada esta topologia.

O erro médio quadrático cometido pela rede auto-associativa n x p x n pode ser escrito como:

$$EQ = E[(\mathbf{x} - \overline{\mathbf{W}}\underline{\mathbf{W}}^{\mathbf{T}}\mathbf{x})^2], \tag{3.29}$$

onde $\underline{\mathbf{W}}$ e $\overline{\mathbf{W}}$ são matrizes com a mesma dimensão (n x p), correspondentes aos pesos

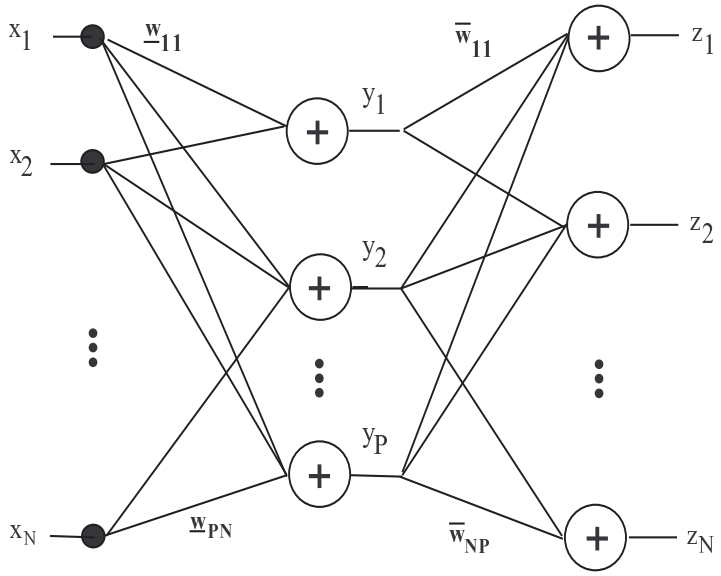


Figura 3.1: Arquitetura neural n x p x n para a extração das componentes

da primeira e segunda camadas da rede, respectivamente. Comparando a Equação 3.29 com a Equação 3.4, pode-se perceber que o treinamento desta rede pode ser explorado para a extração de componentes.

Um dos primeiros trabalhos utilizando a arquitetura auto-associativa para a extração de componentes deve-se a Boulard e Kamp (1989) [84], o qual mostra que, caso a matriz de correlação dos dados seja *full-rank* [70], os valores ótimos dos pesos desta topologia são dados por:

$$\overline{\mathbf{W}}^* = \mathbf{QM} \tag{3.30}$$

$$\underline{\mathbf{W}}^* = \mathbf{M}^{-1}\mathbf{Q}, \tag{3.31}$$

onde $\mathbf{Q}$ está relacionado a decomposição SVD da matriz $\mathbf{R_x}$, isto é: $\mathbf{R_x} = \mathbf{Q}^T\mathbf{\Lambda Q}$ e $\mathbf{M}$ é uma matriz inversível, qualquer, de dimensões p x p. Em outras palavras, a arquitetura proposta não extrai as componentes principais propriamente ditas, mas sim o subespaço a elas associado, uma vez que, experimentalmente, verifica-se que não há convergência para a solução onde $\mathbf{M} = \mathbf{I}$ [65].

Uma interpretação sobre a convergência dos pesos para o subespaço das componentes é a seguinte: ela é devida à simetria da estrutura da rede. Em virtude desta simetria, o treinamento resulta num conjunto de parâmetros tal que a contribuição

de cada neurônio da camada intermediária na formação do vetor de saída seja similar. Caso os pesos convergissem para as componentes propriamente ditas, esta contribuição deveria ser distinta.

Para o treinamento desta rede, o artigo propõe, em contraposição ao uso do algoritmo de *backpropagation*, a utilização de ferramentas numéricas de álgebra linear. Em continuidade a este trabalho, Baldi e Hornik (1989) [85] mostram que a superfície de erro associada a esta arquitetura não possui mínimos locais, apenas pontos de sela.

Um método que possui forte relação com as redes auto-associativas, fato que não é descrito pela literatura, é o de Lei Xu (1993) [75]. Este trabalho propõe uma arquitetura de rede auto-organizável [4] com uma ou múltiplas camadas. É demostrado que, caso esta arquitetura possua uma única camada, com função de ativação linear, é possível realizar a extração das componentes do subespaço. A topologia desta rede, com uma única camada, para a extração de p componentes, é apresentada na Figura 3.2.

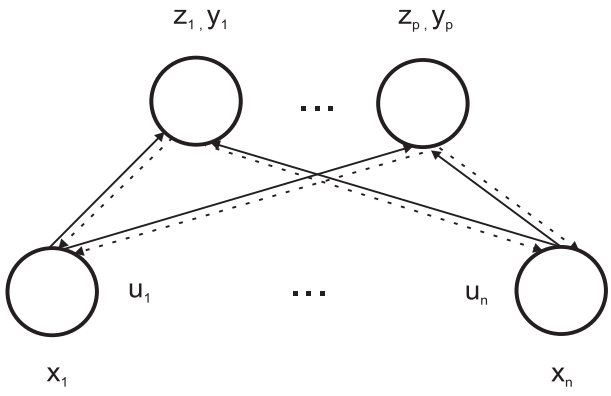


Figura 3.2: Arquitetura da rede auto-organizável proposta por Lei-Xu (uma camada).

Nesta figura é possível identificar: as entradas $(x_1, \ldots, x_n)$, com dimensão n, e a camada neural, que possui p unidades. A dinâmica da rede é a seguinte: o sinal de entrada x_j $(1 \leq j \leq n)$ é propagado das entradas para a camada neural de acordo com a projeção:

$$\mathbf{y} = \mathbf{W}^T \mathbf{x}, \tag{3.32}$$

onde $\mathbf{x}$ é um vetor correspondente às entradas, formado por $x_1 \dots x_n$, e $\mathbf{y}$ reúne os sinais produzidos pelas p-unidades, dados por $y_1 \dots y_p$. A matriz $\mathbf{W}$ é a matriz de pesos, de dimensão n x p, que conecta as entradas à camada neural. Em seguida, é aplicada uma função de ativação ao vetor de sinais $\mathbf{y}$ que produz o vetor $\mathbf{z}$ dado por:

$$\mathbf{z} = S(\mathbf{y}), \tag{3.33}$$

onde S é uma função de ativação qualquer. Por fim, o vetor $\mathbf{z}$ é propagado para a camada de entrada pela equação:

$$\mathbf{u} = \mathbf{W}\mathbf{z}, \tag{3.34}$$

onde a matriz $\mathbf{W}$ é a mesma utilizada pela Equação 3.32, ou seja, a arquitetura utiliza pesos simétricos e bidirecionais.

O treinamento da rede é realizado pela minimização da função objetivo:

$$J = E(|\mathbf{x} - \mathbf{u}|^2), \tag{3.35}$$

que, caso $\mathbf{S}(\mathbf{x}) = \mathbf{x}$, assume a forma:

$$J = E(|\mathbf{x} - \mathbf{W}\mathbf{W}^T\mathbf{x}|^2), \tag{3.36}$$

ou seja, busca-se realizar a reconstrução do vetor de entrada com erro mínimo, o que originou o nome do método: LMSER (*Least Mean Square Error Reconstruction*). Conforme resultados do artigo, a superfície associada a função J não possui mínimos locais, apenas pontos de sela, ou seja, o algoritmo converge globalmente.

Comparando a Equação 3.36 com a Equação 3.29, pode-se perceber que uma rede auto-associativa com pesos idênticos na camada intermediária e de saída possui uma função objetivo àquela da rede LMSER. Desta forma, ambas arquiteturas são equivalentes.

Para produção das equações de treinamento, o artigo realiza a minimização da Equação 3.36 pela técnica de gradiente estocástico, resultando em [75]:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k)[\mathbf{e}(k)\mathbf{y}^T(k) + \mathbf{x}(k)\mathbf{e}(k)^T\mathbf{W}(k)], \tag{3.37}$$

para:

$$\mathbf{e}(k) = \mathbf{x}(k) - \mathbf{W}(k)\mathbf{y}(k) \tag{3.38}$$

$$\mathbf{y}(k) = \mathbf{W}^T(k)\mathbf{x}(k), \tag{3.39}$$

onde $\mathbf{W}$ possui dimensão n x p, ou seja, as componentes do subespaço são extraídas nas colunas de $\mathbf{W}$, e $\eta(k)$ é o fator de aprendizado da rede [4]. Esta arquitetura será aqui referida como LMSER simétrica.

Fato relevante é que as Equações 3.37-3.39 são idênticas às utilizadas no treinamento da topologia MLP auto-associativa, de pesos idênticos, quando é realizada a otimização por gradiente estocástico, o que é conseqüência da similaridade das funções objetivos de ambos métodos, conforme resultados de [86].

Para realizar a extração PCA, o artigo propõe a modificação da função de ativação S de forma que:

$$\mathbf{z} = S(\mathbf{y}) = \mathbf{A}_p\mathbf{y}, \tag{3.40}$$

onde $\mathbf{A}_p$ é uma matriz diagonal com valores positivos e decrescentes, ou seja, foi inserida uma assimetria na estrutura da rede, de forma que a contribuição de cada unidade no erro de reconstrução fosse distinta. Esta arquitetura será aqui referida como LMSER asimétrica.

Para o treinamento, o artigo propõe a minimização da função objetivo dada pela Equação 3.11, derivando, por gradiente estocástico, as seguintes equações de treinamento [75]:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k)[\mathbf{e}(k)\mathbf{y}^T(k)\mathbf{A}_p + \mathbf{x}(k)\mathbf{e}(k)^T\mathbf{W}(k)\mathbf{A}_p], \tag{3.41}$$

onde:

$$\mathbf{e}(k) = \mathbf{x}(k) - \mathbf{W}(k)\mathbf{A}_p\mathbf{y}(k) \tag{3.42}$$

$$\mathbf{y}(k) = \mathbf{W}^T(k)\mathbf{x}(k), \tag{3.43}$$

para o fator de aprendizado $\eta(k)$. Pode-se observar, conforme esperado, que caso $\mathbf{A}_m = \mathbf{I}$, as Equações 3.41-3.43 assumem a forma das Equações 3.37-3.39.

De forma similar, considerando a analogia verificada entre a rede LMSER simétrica e a rede MLP auto-associativa, é possível propor uma arquitetura MLP auto-associativa equivalente à rede LMSER asimétrica. Esta rede deve possuir as seguintes equações de propagação:

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \tag{3.44}$$

$$\mathbf{z} = \mathbf{W}\mathbf{A}_p\mathbf{y}, \tag{3.45}$$

e corresponde a topologia representada na Figura 3.3. Sua principal diferença com respeito à arquitetura MLP usual é a existência de fatores multiplicativos $(a_1, \ldots, a_p)$ nas saídas da camada intermediária, que correspondem aos valores da diagonal da matriz $\mathbf{A}_p$. Esta topologia será aqui referida como rede auto-associativa assimética de pesos idênticos, tendo sido explorada pelo algoritmo WNIC [81].

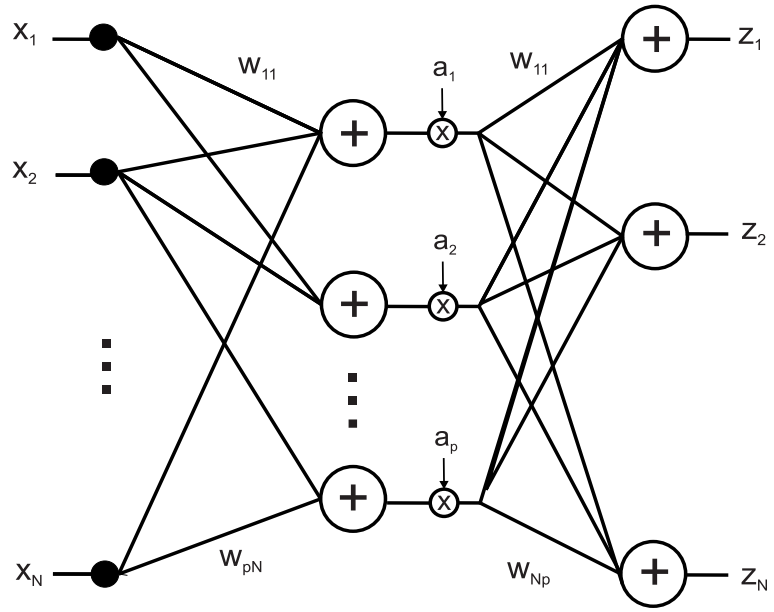


Figura 3.3: Arquitetura neural n x p x n, assimétrica, de pesos idênticos, para a extração de componentes principais

No artigo [87], Diamantaras e Kung (1994) dedicam-se ao estudo de redes lineares MLP auto-associativas, cuja dimensão dos alvos e das entradas é distinta, o que é referido como auto-associador genérico. Esta rede é conhecida como rede hetero-supervisionada [64], sendo uma extensão natural da proposta por Bourlard e Kamp (1989) [84].

Suponha uma rede MLP de dimensões: n x p x m, onde $p \leq \min\{n, m\}$. Sejam $\mathbf{X}$ e $\mathbf{Y}$ matrizes cujas colunas são formadas por N eventos de entrada e alvos da rede, respectivamente. As dimensões de $\mathbf{X}$ e $\mathbf{Y}$ são, portanto, n x N e m x N. O trabalho de Diamantaras e Kung mostra que o treinamento desta rede produz pesos relacionados à decomposição por valor singular generalizada (GSVD) do par de matrizes $\mathbf{Y}\mathbf{X}^T$ e $\mathbf{X}$. A matriz $\mathbf{Y}\mathbf{X}^T$ é conhecida como matriz de correlação cruzada

entre as entradas e os alvos.

Por definição, a decomposição por valor singular generalizada (GSVD) do par de matrizes $\mathbf{YX}^T$ e $\mathbf{X}$ é caracterizada pelas matrizes $\mathbf{U} \in R^{NxN}$, $\mathbf{Q} \in R^{nxn}$ e $\mathbf{V} \in R^{mxm}$, as quais devem satisfazer [88]:

$$\mathbf{U}^T\mathbf{X}^T\mathbf{Q} = \mathbf{D}_1 = diag\{\alpha_1, \ldots, \alpha_n\} \in R^{Nxn} \tag{3.46}$$

$$\mathbf{V}^T\mathbf{Y}\mathbf{X}^T\mathbf{Q} = \mathbf{D}_2 = diag\{\beta_1, \ldots, \beta_n\} \in R^{mxn}, \tag{3.47}$$

onde $\alpha_1 > \ldots > \alpha_r > 0$ e $\alpha_{r+1}, \ldots, \alpha_n = 0$, para algum r, tal que $1 \leq r \leq n$; e $\beta_1 > \ldots > \beta_s > 0$ e $\beta_{s+1}, \ldots, \beta_q \geq 0$, para algum s, tal que $1 \leq s \leq q$. Em outras palavras, as matrizes $\mathbf{U}$ e $\mathbf{Q}$ realizam a diagonalização da matriz de entrada; e as matrizes $\mathbf{V}$ e $\mathbf{Q}$, da matriz de correlação cruzada entre as entradas e os alvos.

A relação entre a rede MLP n x p x m e a GSVD do par ($\mathbf{YX}^T$,$\mathbf{X}$) pode ser estabelecida pelos pesos ótimos desta arquitetura de rede, que são dados por [87]:

$$\overline{\mathbf{W}} = \mathbf{V}_p\mathbf{M} \tag{3.48}$$

$$\underline{\mathbf{W}} = (\mathbf{Q}_p\mathbf{D}_{1p}^{-1} + \mathbf{Q}_1\Phi)\mathbf{\Lambda_p}\mathbf{M}^{-T}, \tag{3.49}$$

onde $\mathbf{\Phi}$ é uma matriz qualquer $(n-r)$ x p, $\mathbf{Q}_p$ e $\mathbf{V}_p$ correspondem a matrizes formadas pela p primeiras colunas de $\mathbf{Q}$ e $\mathbf{V}$, respectivamente; $\mathbf{Q}_1$ é formada pelas $(n-r)$ últimas colunas de $\mathbf{Q}$, $\mathbf{M}$ é uma matriz p x p arbitrária, $\mathbf{D}_{1p}$ é uma matriz diagonal p x p, formada pelos p primeiros elementos de $\mathbf{D}_1$, e $\mathbf{\Lambda}_p$ é também diagonal, de dimensões p x p, com elementos dados por $\frac{\beta_i}{\alpha i}$ (veja as Equações 3.46-3.47). Conclui-se que o peso $\overline{\mathbf{W}}$ converge para um subespaço associado a $\mathbf{V_p}$, enquanto $\underline{\mathbf{W}}$ converge para o subespaço associado a $\mathbf{Q}$. Um caso particular relevante é $\mathbf{Y} = \mathbf{X}$ (auto-associativo), com $\mathbf{X}$ *full-rank*, para o qual os pesos ótimos assumem uma forma idêntica a derivada por Boulard e Kamp [84].

Para o treinamento da estrutura, é proposto o algoritmo de *backpropagation* com gradiente descendente estocástico. Para extender a arquitetura à extração das componentes principais, duas alternativas foram propostas: uma envolvendo múltiplas redes com vetores-alvo submetidos à deflação; e uma outra, explorando conexões laterais entre os neurônios. Para maiores detalhes, consultar a referência [87].

Na Tabela 3.2, apresentamos as principais características dos metodos auto-associativos discutidos. Cabe observar que, para todas as propostas, exceto a de Boulard e Kamp [84], são obtidos algoritmos capazes de operação *on-line*.

Tabela 3.2: Principais características dos algoritmos auto-supervisionados discutidos

Método	**Equação**	**Otimização**	**Referências**	**Extração**
Auto-associativo	-	-	Boulard (1989)[84]	PSA
LMSER Simétrico	(3.37)	Gradiente Estocástico	Lei Xu (1993) [75]	PSA
LMSER Assimétrico	(3.41)	Gradiente Estocástico	Lei Xu (1993) [75]	PCA
Hetero-supervisionado	-	Gradiente Estocástico	Diamantaras (1994) [87]	PSA/PCA

3.4.2 Redes por maximização de variância

Um dos trabalhos pioneiros que relacionam o treinamento de um neurônio linear com a extração de componentes principais deve-se a Amari (1978) [89]. Trata-se de um trabalho pouco conhecido, citado recentemente em [67]. Neste trabalho é mostrado que caso um neurônio seja treinado visando maximizar a variância de sua saída, seus pesos convergem para a direção da componente principal. As equações de treinamento do método de Amari são as seguintes:

$$\tilde{\mathbf{w}}_1(k+1) = \mathbf{w}_1(k) + \eta(k) y_1(k) \mathbf{x}(k) \tag{3.50}$$

$$\mathbf{w}_1(k+1) = \frac{\tilde{\mathbf{w}}_1(k+1)}{||\tilde{\mathbf{w}}_1(k+1)||}, \tag{3.51}$$

para:

$$y_1(k) = \mathbf{w}_1^T(k) \mathbf{x}(k), \tag{3.52}$$

e $\eta(k)$ corresponde ao fator de aprendizado da rede [4].

A interpretação destas equações de treinamento é a seguinte: a Equação 3.50 resulta da maximização, sem restrições, da variância da saida do neurônio $y(k)$, a qual segue o princípio hebbiano [4]. Como a otimização é realizada sem restrições, para evitar o crescimento indiscriminado de $\tilde{\mathbf{w}}_1$, a Equação 3.51 normaliza o vetor de pesos por seu módulo a cada iteração.

No trabalho de Oja (1982) [61], extensivamente citado na literatura, a normalização é incorporada à própria equação de treinamento, resultando na seguinte equação:

$$\mathbf{w}_1(k+1) = \mathbf{w}_1(k) + \eta(k) y_1(k) [\mathbf{x}(k) - y_1(k) \mathbf{w}_1(k)], \tag{3.53}$$

para:

$$y_1(k) \quad = \mathbf{w}_1^T(k)\mathbf{x}(k) \tag{3.54}$$

Vários trabalhos se dedicaram a generalizar o trabalho de Oja para a extração de múltiplas componentes. Um destes trabalhos deve-se a Oja e Karhunen (1985) [90], que propuseram o *Stochastic Gradient Ascent* (SGA), o qual utiliza, para a extração de p componentes, uma rede de p neurônios lineares, conforme a Figura 3.4.

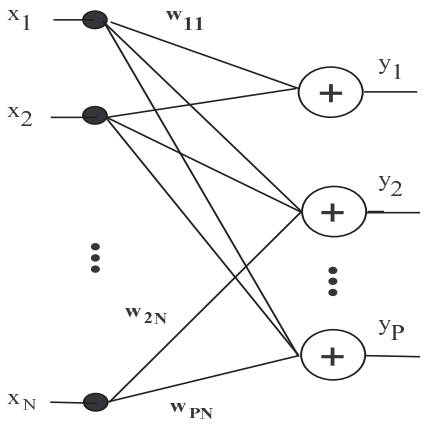


Figura 3.4: Arquitetura para a extração PCA por otimização de variância

No método SGA, para a extração de p componentes, é proposto um aprendizado segundo a regra:

$$\tilde{\mathbf{W}}(k) = \tilde{\mathbf{W}}(k-1) + \gamma(k)\mathbf{x}(k)\mathbf{y}(k)^T \tag{3.55}$$

$$\mathbf{W}(k) = \tilde{\mathbf{W}}(k)\mathbf{S}(k)^{-1} \tag{3.56}$$

para:

$$\mathbf{y}(k) = \mathbf{W}(k-1)^T\mathbf{x}(k), \tag{3.57}$$

onde as dimensões de $\mathbf{x}$, $\mathbf{y}$ e $\mathbf{W}(k)$ são n x 1, p x 1 e n x p, respectivamente, onde n corresponde a dimensão dos dados de entrada. A lógica do algoritmo é a seguinte: na Equação 3.55 tem-se uma maximização de variância, de forma similar à Equação 3.50, enquanto na Equação 3.56 é realizada uma ortonomalização dos pesos da rede, ou seja, o papel da matriz $\mathbf{S}(k)$ é forçar que os pesos de cada neurônio sejam

ortogonais entre si e possuam peso unitário. É possível perceber uma clara relação desta proposta com a estratégia utilizada para o coeficiente de Rayleigh (vide seção 3.3.3). A diferença principal reside na forma como é imposta a normalização: para a derivação por coeficiente de Rayleigh, utilizam-se os multiplicadores de Lagrange; aqui na proposta SGA, utiliza-se a transformação linear $\mathbf{S}(k)$.

O trabalho de Oja e Karhunen [90] mostra que, se $\mathbf{S}(k)$ realizar a ortogonalização de Gram-Schmidt dos pesos da rede a cada iteração, as Equações 3.55-3.57 podem ser reescritas na forma:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)y_j(k)\Big[\mathbf{x}_j(k) - y_j(k)\mathbf{w}_j(k) + \mathbf{p}_j(k)\Big] \tag{3.58}$$

$$y_j(k) = \mathbf{w}_j^T(k)\mathbf{x}(k) \tag{3.59}$$

$$\mathbf{x}_j(k) = \mathbf{x}(k) - \sum_{i=1}^{j-1} y_i(k)\mathbf{w}_i(k) \tag{3.60}$$

$$\mathbf{p}_j(k) = -\sum_{i=1}^{j-1} y_i(k)\mathbf{w}_i(k), \tag{3.61}$$

produzindo um algoritmo de extração PCA. Vale observar que no SGA há uma convergência ordenada para as componentes, ou seja, $\mathbf{w}_1$ converge para a primeira componente, $\mathbf{w}_2$ para a segunda, e assim suscessivamente.

Seguindo esta linha de raciocínio, tem-se o *Subspace Network Learning Algorithm* (SNLA) de Oja (1989) [37]), freqüentemente citado na literatura. No SNLA, o aprendizado também é realizado segundo às Equações 3.55-3.57, e a matriz $\mathbf{S}(k)$ é escolhida visando promover uma ortogonalização simétrica de $\mathbf{W}$ [37], resultando nas seguintes equações de treinamento:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)y_j(k)\mathbf{x}_j(k) \tag{3.62}$$

$$y_j(k) = \mathbf{w}_j^T(k)\mathbf{x}(k) \tag{3.63}$$

$$\mathbf{x}_j(k) = \mathbf{x}(k) - \sum_{i=1}^{p} y_i(k)\mathbf{w}_i(k), \tag{3.64}$$

que convergem para o subespaço das componentes. Caso as Equações 3.62-3.64 sejam reescritas em forma matricial, resulta:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k)\mathbf{e}(k)\mathbf{y}^T(k) \tag{3.65}$$

para:

$$\mathbf{e}(k) = \mathbf{x}(k) - \mathbf{W}(k)\mathbf{y}(k) \tag{3.66}$$

$$\mathbf{y}(k) = \mathbf{W}^T(k)\mathbf{x}(k), \tag{3.67}$$

forma a qual é similar àquela derivada no método LMSER simétrico (Equações 3.37-3.39), exceto pelo termo $\mathbf{x}(k)\mathbf{e}(k)^T\mathbf{W}(k)$. Esta conexão foi estabelecida, primeiramente, por Lei Xu (1993) [75], o qual afirmou que o algoritmo SNLA consiste numa aproximação da técnica LMSER simétrica.

Outro trabalho, freqüentemente citado na literatura, deve-se a Sanger (1989) [38], que propõe o *Generalized Hebbian Algorithm* (GHA). As equações de treinamento do método são:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)y_j(k)[\mathbf{x}_j(k) - y_j(k)\mathbf{w}_j(k)] \tag{3.68}$$

$$y_j(k) = \mathbf{w}_j^T(k)\mathbf{x}(k) \tag{3.69}$$

$$\mathbf{x}_j(k) = \mathbf{x}(k) - \sum_{i=1}^{j-1} y_i(k)\mathbf{w}_j(k), \tag{3.70}$$

e o método converge, ordenadamente, para as componentes. Comparando as Equações 3.68-3.70 com as Equações 3.58-3.61, percebe-se que os métodos são quase idênticos, exceto pelo termo $\mathbf{p}_j(k)$ na Equação 3.58, fato que não é evidenciado pela literatura.

Um outro método é o *Weighted Subspace Algorithm*, que foi proposto por Oja, Ogawa e Wangviwattana (1992) [91, 92], e possui uma relação interessante com o SNLA, também não evidenciada pela literatura. As equações de treinamento do método WSA podem ser escritas na forma:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)y_j(k)\mathbf{x}_j(k) \tag{3.71}$$

$$y_j(k) = \mathbf{w}_j^T(k)\mathbf{x}(k) \tag{3.72}$$

$$\mathbf{x}_j(k) = \mathbf{x}(k) - \theta_j \sum_{i=1}^{p} y_i(k)\mathbf{w}_i(k), \tag{3.73}$$

onde θ_j são constantes positivas decrescentes, escolhidas arbitrariamente. O método WSA realiza a extração das componentes principais. Pode-se perceber que, a menos das constantes, as equações de treinamento dos métodos SNLA e WSA são idênticas. Conclui-se que, no método WSA, é realizada a maximização da variância das saídas dos neurônios, cada um deles contribuindo de forma diferenciada, determinada

pelas constantes θ_i. Vale observar que esta estratégia é similiar àquela adotada por Brockett (vide seção 3.3.1).

Uma outra relação pode ser estabelecida entre o método WSA e o LMSER assimétrico. Reescrevendo as Equações 3.71-3.73 na forma matricial, resulta:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k)\mathbf{e}(k)\mathbf{y}^T(k)\mathbf{A}_p \tag{3.74}$$

para:

$$\mathbf{e}(k) = \mathbf{x}(k) - \mathbf{W}(k)\mathbf{A}_p\mathbf{y}(k) \tag{3.75}$$

$$\mathbf{y}(k) = \mathbf{W}^T(k)\mathbf{x}(k), \tag{3.76}$$

onde $\mathbf{A}_p$ é uma matriz p x p, cuja diagonal é formada pelos valores de θ_j. Pode-ser observar que a Equação 3.74 é similar a Equação 3.41, exceto pelo termo $\mathbf{x}(k)\mathbf{e}(k)^T\mathbf{W}(k)\mathbf{A}_m$. De forma similar à relação estabelecida, anteriormente, entre os métodos SNLA e LMSER-simétrico, conclui-se que o algoritmo WSA consiste numa aproximação do algoritmo LMSER-assimétrico, o que não é mencionado pela literatura.

3.4.2.1 Métodos de inibição lateral

Uma outra estratégia, na linha de maximização de variância, consiste em explorar as inibições laterais. Esta arquitetura foi proposta originalmente por Rubner e Tavan (1989) [93], sendo apresentada na Figura 3.5.

Da figura é possível perceber que, adicionalmente aos pesos convencionais ($\mathbf{w}_j$), no qual as componentes são extraídas, são estabelecidas, de forma hierárquica, conexões laterais entre os neurônios ($\mathbf{c}_j$), cujo papel é realizar a ortogonalização dos pesos da rede.

Um método de destaque na literatura que explora as conexões laterais é o *Adaptive Principal Component Extraction* (APEX), proposto por Kung, Diamantaras e Taur (1994) [62], cujas equações de treinamento são dadas por:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)y_j(k)[\mathbf{x}(k) - y_j(k)\mathbf{w}_j(k)] \tag{3.77}$$

$$\mathbf{c}_j(k+1) = \mathbf{c}_j(k) + \eta(k)y_j(k)[\mathbf{d}_j(k) - y_j(k)\mathbf{c}_j(k)] \tag{3.78}$$

para:

$$\mathbf{d}_j(k) = [\mathbf{w}_1(k) \quad \ldots \quad \mathbf{w}_{j-1}(k)]^T\mathbf{x}(k) \tag{3.79}$$

$$y_j(k) = \mathbf{w}_j(k)^T\mathbf{x}(k) - \mathbf{c}_j^T(k)\mathbf{d}_j(k), \tag{3.80}$$

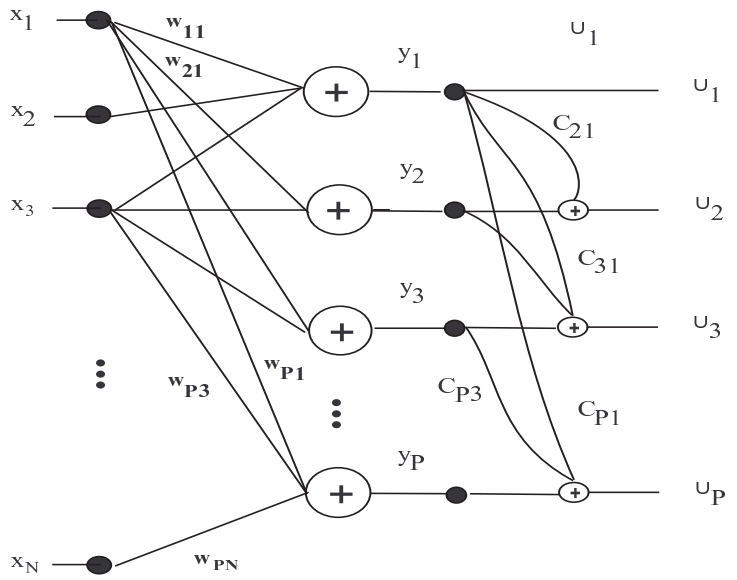


Figura 3.5: Arquitetura de Rubner e Tavan para a extração de componentes principais

onde $\mathbf{w}_j(k)$ responde pelos pesos associados às conexões diretas e $\mathbf{c}_j(k)$ pelas inibições laterais.

Uma das principais vantagens do método APEX é o fato que uma relação com filtros adaptativos pode ser estabelecida, a qual permite que o valor de $\eta(k)$ seja determinado, de forma ótima, a cada iteração. Este cálculo produz um impacto significativo na convergência do algoritmo [62]. A equação iterativa para a determinação de $\eta(k)$ é dada por [62]:

$$\eta(k) = \frac{\eta(k-1)}{\beta + \eta(k-1)y_j^2(k)}, \tag{3.81}$$

onde β é o fator de esquecimento, que deve assumir um valor menor ou igual a um, de forma similar à função objetivo PAST (vide seção 3.3.6).

Algumas extensões foram desenvolvidas sobre o método APEX, entre elas uma devida a Fiori e Piazza (2000) [94] que, aplicando critérios alternativos de otimização, propõem variantes do método APEX original, referidas com ψ-APEX, as quais apresentaram uma melhor convergência nas simulações apresentadas.

Na Tabela 3.3 resumimos os diferentes algoritmos de extração das componentes baseados em maximização de variância. Todos algoritmos discutidos realizam

extração *on-line*.

Tabela 3.3: Principais características dos algoritmos baseados em variância

Nome	**Equações**	**Otimização**	**Trabalhos**	**Mod.**
Regra de Amari	(3.50-3.52)	Gradiente Estocástico	Amari (1978) [89]	$1^{a.}$ PCA
Regra de Oja	(3.53-3.54)	Gradiente Estocástico	Oja (1982) [61]	$1^{a.}$ PCA
SGA	(3.58-3.61)	Gradiente Estocástico	Oja (1985) [90]	PCA
SNLA	(3.62-3.64)	Gradiente Estocástico	Oja (1989) [37]	PSA
GHA	(3.68-3.70)	Gradiente Estocástico	Sanger (1989) [38]	PCA
WSA	(3.71-3.73)	Gradiente Estocástico	Oja (1992) [91, 92]	PCA
APEX	(3.77-3.80)	Gradiente Estocástico/RLS	Kung (1994) [62]	PCA

3.5 Algoritmos de convergência otimizada

Grande parte dos algoritmos discutidos anteriormente, por basear-se na técnica de gradiente descendente, apresenta uma velocidade de convergência relacionada à escolha da constante de aprendizado. Dada a simplicidade desta técnica, a convergência é normalmente lenta, em especial para aplicações com dimensão elevada e grande número de eventos. Este fato motivou a pesquisa na literatura por algoritmos mais eficientes para a extração PCA. Este algoritmos serão apresentados a seguir.

3.5.1 NIC e WNIC

Os algoritmos NIC e WNIC, cujas funções objetivo foram anteriormente citadas, baseiam-se numa arquitetura neural de duas camadas, idêntica a rede auto-associativa simétrica (NIC) e assimétrica (WNIC) de pesos idênticos (vide seção 3.4.1). Conforme já discutido, o algoritmo NIC [80] realiza a extração das componentes do subespaço, e o WNIC [81] extrai as componentes principais. Duas versões destas técnicas foram propostas: uma por batelada, baseada em gradiente descendente; e outra, que utiliza uma função recursiva (RLS). A primeira versão é dependende da estimação da matriz de correlação dos dados e destina-se a operação *off-line*, enquanto a segunda opera *on-line* e não exige esta estimação. Vale destacar

que para o algoritmo WNIC, a acuidade das componentes extraídas é depende da escolha de um conjunto de constantes decrescentes arbitrárias [81].

3.5.2 Chatterjee, Kang e Roychowdhury

Estes autores propõem em [78] uma série de algoritmos rápidos baseados na otimização da função objetivo determinada pela Equação 3.18, realizada através das técnicas de gradiente descendente, *steepest descent*, gradiente conjugado e Newton-Rapson. Um dos principais inconvenientes das técnicas propostas é exigir a estimação da matriz de correlação dos dados e o alto-custo computacional das equações de extração propostas. Tratam-se de algoritmos *off-line*.

3.5.3 Redes RLS

O algortimo de Bannour e Azimi-Sadjadi (1995) [82], aqui referido como BS-RLS, utiliza p-redes MLP auto-associativas, de dimensões n x 1 x n, com pesos da primeira e segunda camadas iguais. A atualização dos pesos de cada rede é realizada pela otimização da função objetivo apresentada na Equação 3.24, resultando nas seguintes equações de treinamento [82]:

$$h_j(n) = \mathbf{w}_j^T(n-1)\mathbf{x}(n) \tag{3.82}$$

$$K_j(n) = \frac{P_j(n-1)h_j(n)}{[1 + h_j(n)^2 P_j(n-1)]} \tag{3.83}$$

$$\mathbf{w}_j(n) = \mathbf{w}_j(n-1) + K_j(n)[\mathbf{d}_j(n) - h_j(n)\mathbf{w}_j(n-1)] \tag{3.84}$$

$$P_j(n) = [1 - K_j(n)h_j(n)]P_j(n-1), \tag{3.85}$$

para $\mathbf{d}_j(n)$ dado pela Equação 3.23.

Outro algoritmo nesta linha, que não assume explicitamente nenhuma arquitetura neural, é o PAST, proposto por Yang (1995) [69]. Neste trabalho é apresentada, sem nenhuma derivação, uma versão para a extração das componentes principais, referida como PASTd. As equações de treinamento do método PASTd

são dadas por:

$$\text{Repetir } t = 1, 2, \ldots$$

$$\mathbf{x}_1(t) = \mathbf{x}(t) \tag{3.86}$$

$$\text{Repetir } i = 1, 2, \ldots, p$$

$$y_i(t) = \mathbf{w}_i^T(t-1)\mathbf{x}_i(t) \tag{3.87}$$

$$d_i(t) = \beta d_i(t-1) + |y_i(t)|^2 \tag{3.88}$$

$$\mathbf{e}_i(t) = \mathbf{x}_i(t) - \mathbf{w}_i(t-1)y_i(t) \tag{3.89}$$

$$\mathbf{w}_i(t) = \mathbf{w}_i(t-1) + \frac{1}{d_i(t)} y_i(t)\mathbf{e}_i(t) \tag{3.90}$$

$$\mathbf{x}_{i+1}(t) = \mathbf{x}_i(t) - \mathbf{w}_i(t)y_i(t), \tag{3.91}$$

onde $\mathbf{x}(t)$ corresponde ao dado de entrada da t-ésima iteração, β é o fator de esquecimento (vide seção 3.3.6) e p é o número de componentes a extrair.

Um método que possui uma forte relação com o algoritmo PASTd, o que não é evidenciado pela literatura, deve-se a Cichocki, Kasprzak e Skarbek (1996) [95]. Neste método, para a extração de p componentes, são utilizadas p redes neurais, as quais são dispostas em cascata. Esta arquitetura possui equações de treinamento idênticas a do método PASTd, podendo ser considerada uma implementação neural desta técnica.

3.5.4 Método das potências

Um metodo interessante, derivado no ambiente de álgebra linear, é o método das potências [88]. Este método é baseado na seguinte equação recursiva:

$$\mathbf{w}_1(k+1) = \mathbf{A}\mathbf{w}_1(k), \tag{3.92}$$

onde $\mathbf{A}$ é uma matriz arbitrária. Para a extração das componentes principais, considera-se $\mathbf{A} = \mathbf{R}_\mathbf{x}$, onde $\mathbf{R}_\mathbf{x}$ é a matriz de correlação dos dados.

Sejam $\mathbf{e}_1, \ldots, \mathbf{e}_n$ os autovetores associados aos autovalores $\lambda_1 \ldots \lambda_n$ da matriz $\mathbf{A}$, tal que $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. Se considerarmos um vetor inicial dado por:

$$\mathbf{w}(0) = a_1\mathbf{e}_1 + \ldots + a_n\mathbf{e}_n, \tag{3.93}$$

a Equação 3.92 pode ser reescrita como [88]:

$$\mathbf{w}_1(k+1) = a_1\lambda_1^k\left(\mathbf{e}_1 + \sum_{i=2}^{n} \frac{a_i}{a_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{e}_i\right), \tag{3.94}$$

ou seja, o método converge para uma versão escalada do autovetor dominante de $\mathbf{A}$ (componente principal de $\mathbf{x}$), a uma taxa determinada pela razão $(\frac{\lambda_2}{\lambda_1})$, a qual está associada à exponencial de decaimento mais lento.

Para extender o método à extração da j-ésima componente, vamos considerar a seguinte equação recursiva:

$$\mathbf{w}_j(k+1) = \mathbf{R}_{\mathbf{x}_j}\mathbf{w}_j(k), \tag{3.95}$$

onde a matriz $\mathbf{R}_{\mathbf{x}_j}$ é dada por:

$$\mathbf{R}_{\mathbf{x}_j} = \mathbf{T}_j\mathbf{R}_{\mathbf{x}} \tag{3.96}$$

e

$$\mathbf{T}_j = \left(\mathbf{I} - \sum_{i=1}^{j-1}\mathbf{e}_i\mathbf{e}_i^T\right) \tag{3.97}$$

A matriz $\mathbf{T}_j$ é uma transformação linear conhecida como transformação de deflação, cujo efeito sobre $\mathbf{R}_{\mathbf{x}}$ é zerar seus $(j-1)$ autovalores dominantes [64]. Em outras palavras, a transformação de deflação torna, principal, o j-ésimo autovetor dominante de $\mathbf{R}_{\mathbf{x}}$.

Para analisar a convergência da Equação 3.95, basta notar que a Equação 3.94 pode ser reescrita na forma:

$$\mathbf{w}_j(k+1) \quad = \sum_{i=1}^{j-1} a_i\lambda_i^k\mathbf{e}_i + a_j\lambda_j^k\left(\mathbf{e}_j + \sum_{i=j+1}^{n}\frac{a_i}{a_j}\left(\frac{\lambda_i}{\lambda_j}\right)^k\mathbf{e}_i\right) \tag{3.98}$$

Como os autovetores $\lambda_1 \ldots \lambda_{j-1}$ de $\mathbf{R}_{\mathbf{x}j}$ são nulos por efeito da transformação $\mathbf{T}_j$, a Equação 3.98 assume a forma:

$$\mathbf{w}_j(k+1) = a_j\lambda_j^k\left(\mathbf{e}_j + \sum_{i=j+1}^{n}\frac{a_i}{a_j}\left(\frac{\lambda_i}{\lambda_j}\right)^k\mathbf{e}_i\right), \tag{3.99}$$

a qual converge para uma versão escalada de $\mathbf{e}_j$, a uma taxa ditada pela razão $(\frac{\lambda_j}{\lambda_{j+1}})$.

Em virtude do termo $a_1\lambda_1^k$ na Equação 3.94, e $a_j\lambda_j^k$ na Equação 3.99, para evitar o crescimento infinito do módulo de $\mathbf{w}_j(k)$, deve-se realizar uma normalização. Um critério comumente utilizado consiste em, a cada iteração, normalizar o valor de $\mathbf{w}_j(k+1)$ por seu próprio módulo, produzindo as seguintes equações iterativas:

$$\tilde{\mathbf{w}}_j(k+1) = \mathbf{R}_{\mathbf{x}_j}\mathbf{w}_j(k) \tag{3.100}$$

$$\mathbf{w}_j(k+1) = \frac{\tilde{\mathbf{w}}_j(k+1)}{||\tilde{\mathbf{w}}_j(k+1)||_2}, \qquad j = 1 \ldots p, \tag{3.101}$$

para:

$$\mathbf{R}_{\mathbf{x}_j} = \mathbf{R}_{\mathbf{x}}, \qquad j = 1, \tag{3.102}$$

$$\mathbf{R}_{\mathbf{x}_j} = \left(\mathbf{I} - \sum_{i=1}^{j-1} \mathbf{e}_i \mathbf{e}_i^T\right)\mathbf{R}_{\mathbf{x}}, \qquad j \geq 2. \tag{3.103}$$

O método das potências trata-se, portando, de um processo de extração seqüencial, ou seja, a primeira componente é extraída; logo após, calcula-se $\mathbf{R}_{\mathbf{x}_2}$, considerando $\mathbf{e}_1 \approx \mathbf{w}_1$, e assim, suscessivamente, até a p-ésima componente.

Um inconveniente da forma apresentada pelas Equações 3.100-3.103 é a dependência da matriz de correlação dos dados. É possível, no entanto, estimar $\mathbf{R}_{\mathbf{x}_j}$ por [49]:

$$\mathbf{R}_{\mathbf{x}_j} = k \sum_{i=1}^{M} \mathbf{x}_j(i)\mathbf{x}_j(i)^T, \tag{3.104}$$

onde:

$$\mathbf{x}_j(i) = \mathbf{x}(i) - \sum_{t=1}^{j-1} \mathbf{e}_t \mathbf{e}_t^T \mathbf{x}(i), \tag{3.105}$$

e $k = \frac{1}{M}$, onde M é o número de eventos envolvidos. Outra aproximação útil é considerar válida a aproximação: $\mathbf{w}_i(k) \approx \mathbf{e}_i$, ou seja, que na extração da j-ésima componente, os vetores das $(j-1)$ componentes anteriores convergiram acuradamente. Neste caso, a equação de treinamento da técnica resume-se a [67]:

$$\mathbf{w}_j(k+1) = \frac{\sum_{i=1}^{M} y_j(k,i)\mathbf{x}_j(i)}{||\sum_{i=1}^{M} y_j(k,i)\mathbf{x}_j(i)||} \tag{3.106}$$

para:

$$y_j(k,i) = \mathbf{w}_j(k)^T \mathbf{x}_j(i) \tag{3.107}$$

$$\mathbf{x}_j(i) = \mathbf{x}(i), \qquad j = 1, \tag{3.108}$$

$$\mathbf{x}_j(i) = \mathbf{x}(i) - \sum_{i=1}^{j-1} y_i(k,i)\mathbf{w}_j(k), \qquad j \geq 2, \tag{3.109}$$

cuja vantagem é não exigir o cálculo da matriz de correlação dos dados e de suas versões defletidas, o que provoca sensível redução do custo computacional e das exigências quanto a capacidade de armazenamento. Isto é de extrema valia para conjuntos de dados extensos e com grande número de variáveis. Outra vantagem da forma proposta pelas Equações 3.106-3.109 é permitir que o processo de extração seja realizado de forma paralela, ainda que a convergência seja seqüêncial, dada a natureza do algoritmo. O método das potências é um algoritmo *off-line*.

3.6 Seleção do método

De posse dos vários métodos de extração de componentes principais apresentados, fica a pergunta: qual escolher ? Serão realizadas, a seguir, algumas considerações para orientar esta seleção.

Para os métodos que dependem da escolha de constantes arbitrárias, como por exemplo: LMSER-assimétrico, WNIC e WSA, a literatura reporta que a acuidade das componentes extraídas é dependente do valor das constantes escolhidas. Simulações desenvolvidas pelo autor em conjuntos sintéticos e reais confirmaram esta afirmativa. Em relação ao método hetero-supervisionado, descrito na seção 3.4.1, simulações com dados de sonar [86, 96, 97] mostram que este método fornece duas estimativas para as componentes, cada uma associada a uma camada da rede, de diferentes acuidades, e sua convergência é lenta.

Entre os métodos baseados em maximização de variância, o APEX apresenta, conforme simulações realizadas em [97], uma menor acurácia e velocidade de convergência que o GHA. Quanto aos métodos GHA e SGA, dada a similaridade entre suas equações de treinamento, o comportamento dos dois métodos é similar. Dos métodos considerados nesta discussão inicial, pelo compromisso entre acuidade, simplicidade e velocidade de extração, recomenda-se, portanto, o GHA, que é um método *on-line*.

Em relação aos algoritmos de convergência otimizada com aplicação *on-line*, destacam-se o PASTd e o BS-RLS. A literatura reporta que para processos não-estacionários, a constante $K_j(n)$ (Equação 3.83) deve ser reinicializada periodicamente [82], logo a escolha recai sobre o método PASTd. Quando comparados os métodos PASTd e GHA, a literatura [66] reporta que a velocidade de convergência do primeiro é significativamente maior que do segundo, ainda que, em termos de acuidade, o desempenho de ambos seja equivalente, desde que o fator de esquecimento (constante β) seja escolhido de forma apropriada.

Ainda que métodos *on-line* possam ser utilizados em aplicações *off-line*, a aplicação de algoritmos *off-line* para estas aplicações resulta, normalmente, numa maior velocidade de convergência [65]. Dentre os algoritmos apresentados para aplicações *off-line*, destaca-se o método das potências, que conjuga simplicidade e velocidade de convergência. Em [97], o método das potências apresentou uma maior acurácia e velocidade de convergência em relação ao método PASTd na extração de

componentes de dados de sonar. Recomenda-se, portanto, o método das potências para aplicações que exijam uma extração *off-line* de componentes, em especial, para aquelas com alta dimensionalidade e grande número de eventos.

3.7 Extração de componentes principais por hierarquia de células auto-associativas

Na análise dos métodos anteriores, verificou-se que o problema de extração da componentes principais através de redes neurais compreende 3 estágios: a definição de uma arquitetura de extração, a escolha de uma função objetivo e, por fim, a seleção de um algoritmo de otimização.

Número significativo dos métodos discutidos anteriormente se baseia na otimização por gradiente descendente estocástico, possuindo uma convergência bastante lenta, em especial, para aplicações com grande volume de dados e alta-dimensionalidade. Em [97], utilizando um conjunto de dados de sonar passivo, com um número mais restrito de classes para os contatos (um total de 4 classes e 33 navios), verificou-se que a escolha do fator de aprendizado possui impacto significativo, tanto na velocidade de convergência, quanto na acuidade das componentes extraídas. Foi constatado ainda que técnicas onde o ajuste do fator de aprendizado é realizado de forma automática (PASTd [69], por exemplo) possuiam uma maior velocidade de convergência e melhor acuidade, mostrando-se capaz de extrair, inclusive, um maior número de componentes com a acuidade desejada.

Tais fatores motivaram o desenvolvimento de uma nova sistemática de extração, a qual é baseada numa arquitetura neural específica que é submetida a diferentes algoritmos de otimização. Como a arquitetura é definida de forma independente do algoritmo de otimização, é possível derivar uma família de algoritmos, cada qual mais adequado a uma aplicação particular. Através do modelo proposto é possível, também, obter e generalizar alguns métodos da literatura, o que é atraente, inclusive para a comparação, escolha e implementação das diferentes técnicas em situações práticas, nas quais a dimensionalidade e as restrições de armazenamento impõem sanções em muitos dos algoritmos presentes na literatura.

Em [86] foi mostrado que é possível realizar a extração PCA através de redes

auto-associativas lineares de duas camadas submetidas a um processo de treinamento incremental e construtivo. Segundo este processo, o treinamento da rede é inicializado com apenas um neurônio. Após a convergência, mais um neurônio é inserido e os pesos associados aos demais neurônios são mantidos "congelados"pelo resto de todo o treinamento. O processo é repetido até a convergência do p-ésimo neurônio, onde p corresponde ao número de componentes que se deseja extrair. Duas topologias foram consideradas: a primeira, que considera o treinamento independente dos pesos da primeira e segunda camadas; e a segunda, na qual os pesos de ambas camadas, considerados idênticos, são atualizados simultâneamente. Para o enfoque que utiliza pesos distintos, foi verificada uma diferença significativa de acuidade entre as componentes extraídas pela primeira e segunda camadas, o que prejudica a convergência do algoritmo. Pela maior acurácia de extração e menor número de iterações para convergência, a arquitetura com pesos idênticos mostrou-se, portanto, mais adequada para a extração das componentes. Ambas sistemáticas possuem forte conexão com as propostas de Boulard e Kamp [98] (pesos distintos) e LMSER simétrico [75] (pesos iguais), realizando, no entanto, a extração de componentes principais. Em [96] é proposta uma forma automática para o ajuste do fator de aprendizado desta topologia construtiva, que mostrou produzir significativo ganho na velocidade de convergência.

A nova proposta consiste numa evolução dos trabalhos [86, 96] e utiliza uma rede MLP auto-associativa linear de 2 camadas, com dimensões n x 1 x n, e pesos idênticos, conforme a Figura 3.6. Esta rede será aqui referida como célula auto-associativa. Pela relação direta com o método LMSER-simético [75] (vide seção 3.4.1), concluímos que a célula auto-associativa extrai a componente principal dos dados.

Para generalizar a aplicação da célula auto-associativa permitindo a extração de uma componente principal qualquer, uma estratégia possível é reduzir o problema da extração de p componentes a p problemas independentes mais simples, onde há sempre a extração da primeira componente principal. Uma opção é aplicar uma transformação linear sobre os dados que altere sua componente principal. Em outras palavras, para a extração da segunda componente, a célula opera sobre dados transformados, sendo esta transformação escolhida de forma a eliminar dos

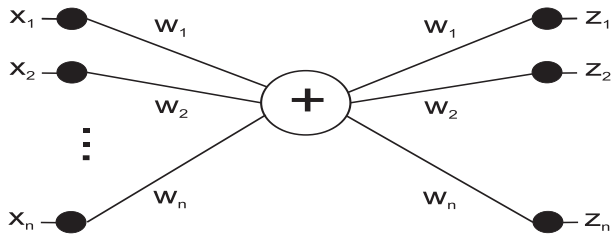


Figura 3.6: Célula auto-associativa

dados todas suas projeções na direção da primeira componente, o que torna a segunda componente a componente principal[2]. Para a terceira, utiliza-se procedimento idêntico, e a transformação elimina a primeira e a segunda componentes; e assim, suscessivamente. A transformação que elimina uma ou mais componentes principais é a transformação de deflação [64]. Assim, para a extração da j-ésima componente principal, aplica-se a transformação de deflação sobre os dados, que resulta em:

$$\mathbf{x}_j = \mathbf{x} - \sum_{i=1}^{j-1} \mathbf{e}_i \mathbf{e}_i^T \mathbf{x}, \tag{3.110}$$

onde $\mathbf{x}_j$ representa o vetor de dados modificado, cuja componente principal corresponde a j-ésima componente do vetor original $\mathbf{x}$.

De forma similar ao método das potências, uma forma mais conveniente para a transformação de deflação é:

$$\mathbf{x}_j = \mathbf{x} - \sum_{i=1}^{j-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x}, \tag{3.111}$$

onde $\mathbf{w}_j$ corresponde aos pesos da j-ésima célula. A Equação 3.111 é uma aproximação da Equação 3.110, considerando $\mathbf{w}_i \approx \mathbf{e}_i$, para $1 \leq i < j$, ou seja, é suposto que na extração da j-ésima componente, os pesos das células associadas às componentes anteriores fornecem uma estimativa acurada dos autovetores a elas associados. A estrutura correspondente a Equação 3.111 é apresentada na Figura 3.7.

Um dos atrativos da modalidade proposta é que uma mesma arquitetura é utilizada para extração de qualquer componente. O que determina qual componente será extraída é a definição sobre qual vetor de dados modificado a arquitetura opera.

[2]A estratégia aqui adotada é análoga a explorada na dedução do método das potências (seção 3.5.4)

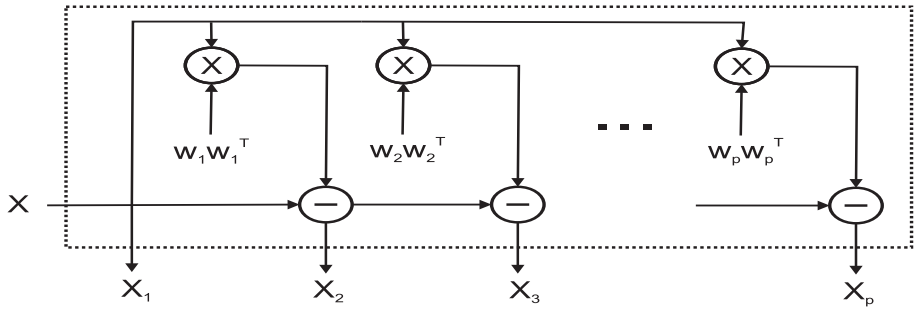


Figura 3.7: Estrutura para a produção de dados defletidos

A estrutura proposta é ainda bastante flexível, permitindo a extração de componentes de forma incremental/seqüencial, onde uma única componente é obtida de cada vez, ou mesmo de forma paralela, onde várias componentes são extraídas "simultaneamente"[3].

Assim, pela sistemática proposta, para a extração de p componentes, utiliza-se uma hierarquia de p células auto-associativas, cada uma submetida às entradas defletidas de modo apropriado. O mesmo resultado pode ainda ser obtido se a hierarquia for submetida a vetores-alvo defletidos. Por fim, é possível ainda aplicar a deflação às entradas e aos alvos, simultaneamente.

Em resumo, considerando $\mathbf{i}_j$ e $\mathbf{t}_j$ as entradas e os vetores-alvo da j-ésima célula auto-associativa, e $\mathbf{x}_j$ dado pela Equação 3.111, há três modalidades de treinamento para a hierarquia de extração proposta:

- **Deflação aplicada às entradas:**

$$\mathbf{i}_j = \mathbf{x}_j \tag{3.112}$$

$$\mathbf{t}_j = \mathbf{x} \tag{3.113}$$

- **Deflação aplicada aos vetores-alvo:**

$$\mathbf{i}_j = \mathbf{x} \tag{3.114}$$

$$\mathbf{t}_j = \mathbf{x}_j \tag{3.115}$$

[3]Vale observar que, a despeito do treinamento dos neurônios ser realizado em paralelo, a convergência das componentes ocorre de forma seqüêncial, em virtude da Equação 3.111

- **Deflação aplicada às entradas e aos vetores-alvo:**

$$\mathbf{i}_j = \mathbf{x}_j \tag{3.116}$$

$$\mathbf{t}_j = \mathbf{x}_j, \tag{3.117}$$

para $1 \leq j \leq p$, onde p corresponde ao número de componentes que se deseja extrair. Estas três modalidades são ilustadas na Figura 3.8.

Cabe mostrar que os pesos ótimos de cada célula auto-associativa simétrica correspondem às componentes principais do processo $\mathbf{x}$. Vale observar que o erro médio quadrático cometido pela j-ésima célula é dado por:

$$K(\mathbf{w}_j(k)) = E\{||\mathbf{t}_j - \mathbf{w}_j\mathbf{w}_j^T\mathbf{i}_j\ ||^2\} \tag{3.118}$$

Utilizando $||\mathbf{a}||^2 = tr\{\mathbf{a}\mathbf{a}^T\}$ e explorando o fato de que $tr\{\mathbf{AB}\} = tr\{\mathbf{BA}\}$ e $tr\{\mathbf{A}^T\} = tr\{\mathbf{A}\}$, é possível reescrever a Equação 3.118 na forma:

$$K(\mathbf{w}_j(k)) = tr\{E[\mathbf{t}_j\mathbf{t}_j^T]\} - 2\mathbf{w}_j^T E[\mathbf{i}_j\mathbf{t}_j^T]\mathbf{w}_j + |\mathbf{w}_j|^2\mathbf{w}_j^T E[\mathbf{i}_j\mathbf{i}_j^T]\mathbf{w}_j \tag{3.119}$$

Se considerarmos que as entradas aplicadas à rede foram defletidas para a extração da r-ésima componente ($\mathbf{i}_j = \mathbf{x}_r$), e os vetores-alvo, para a s-ésima ($\mathbf{t}_j = \mathbf{x}_s$), onde $1 \leq (r, s) \leq j$, tem-se que:

$$tr\{E[\mathbf{t}_j\mathbf{t}_j^T]\} = tr\{\mathbf{R}_{\mathbf{x}_s}\} \tag{3.120}$$

$$\begin{aligned}
E[\mathbf{i}_j\mathbf{t}_j^T] &= E[\mathbf{x}_r\mathbf{x}_s^T] = \left(\mathbf{I}_{nn} - \mathbf{Q}_{(r-1)}\mathbf{Q}_{(r-1)}^T\right)\mathbf{R}_{\mathbf{x}}\left(\mathbf{I}_{nn} - \mathbf{Q}_{(s-1)}\mathbf{Q}_{(s-1)}^T\right)^T \\
E[\mathbf{i}_j\mathbf{t}_j^T] &= \left(\mathbf{I}_{nn} - \mathbf{Q}_{(r-1)}\mathbf{Q}_{(r-1)}^T\right)\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\left(\mathbf{I}_{nn} - \mathbf{Q}_{(s-1)}.\mathbf{Q}_{(s-1)}^T\right)^T \\
E[\mathbf{i}_j\mathbf{t}_j^T] &= \left(\mathbf{Q} - \mathbf{Q}_{(r-1)}\begin{bmatrix}\mathbf{I}_{n(r-1)} & \mathbf{0}_{n(n-r)}\end{bmatrix}\right)\mathbf{\Lambda} \\
&\qquad \left(\mathbf{Q} - \mathbf{Q}_{(s-1)}\begin{bmatrix}\mathbf{I}_{n(s-1)} & \mathbf{0}_{n(n-s)}\end{bmatrix}\right)^T \\
E[\mathbf{i}_j\mathbf{t}_j^T] &= \mathbf{Q}\begin{bmatrix}\mathbf{0}_{n(r-1)} & \mathbf{I}_{n(n-r)}\end{bmatrix}\mathbf{\Lambda}\begin{bmatrix}\mathbf{0}_{n(s-1)} & \mathbf{I}_{n(n-s)}\end{bmatrix}^T\mathbf{Q}^T \\
E[\mathbf{i}_j\mathbf{t}_j^T] &= \mathbf{Q}\text{diag}\{0, \ldots, 0, \lambda_k, \ldots, \lambda_n\}\mathbf{Q}^T \\
E[\mathbf{i}_j\mathbf{t}_j^T] &= \mathbf{R}_{\mathbf{x}_k},
\end{aligned} \tag{3.121}$$

onde $k = \max(r, s)$, *diag* representa uma matriz diagonal, de dimensões n x n, $\mathbf{Q}_j$ é uma matriz cujas colunas são formadas pelos j-autovetores dominantes ordenados

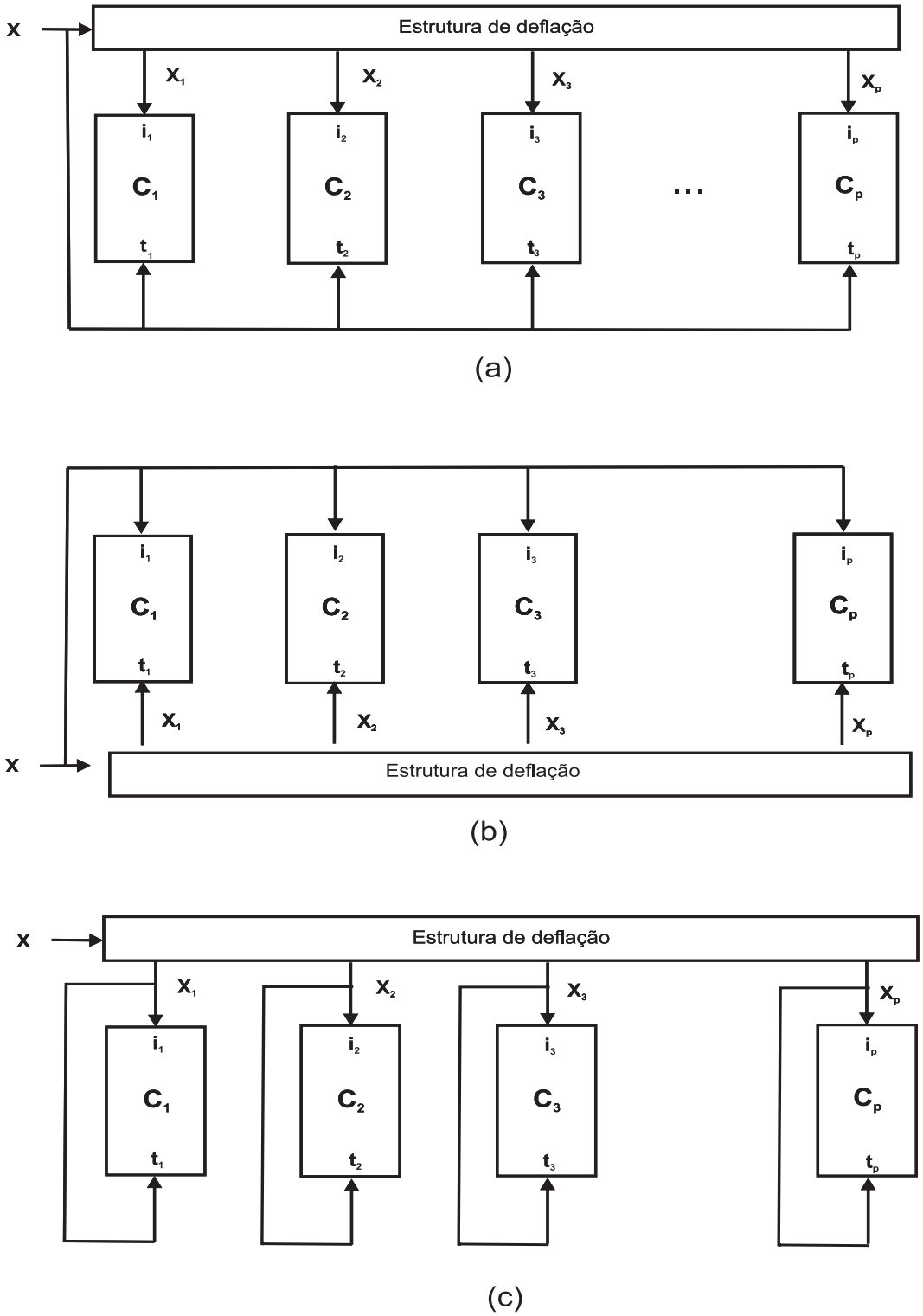


Figura 3.8: Modalidades de extração PCA para a arquitetura proposta: (a) deflação às entradas, (b) deflação aos alvos e (c) deflação às entradas e alvos.

de $\mathbf{Q}$, $\mathbf{I}_{jk}$ é uma matriz retangular, de dimensões j x k , com diagonal unitária e demais elementos nulos, e $\mathbf{0}_{jk}$ é uma matriz de zeros, de dimensões j x k. Conclui-se que a matriz de correlação cruzada entre as entradas e os vetores alvos $E[\mathbf{i}_j\mathbf{t}_j^T]$ possui o autovalor(autovetor) dominante determinado pela deflação de ordem mais alta.

Para as três modalidades de deflação consideradas, o par (r, s) vale $(j, 1)$ (entradas), $(1, j)$ (alvos) e (j, j) (alvos e entradas). Conclui-se pela Equação 3.121 que nos três casos: $k = j$, logo a Equação 3.119 assume a forma:

$$K(\mathbf{w}_j(k)) = tr\{\mathbf{R}_{\mathbf{x}_s}\} - 2\mathbf{w}_j^T\mathbf{R}_{\mathbf{x}_j}\mathbf{w}_j + |\mathbf{w}_j|^2\mathbf{w}_j^T\mathbf{R}_{\mathbf{x}_r}\mathbf{w}_j \tag{3.122}$$

Derivando $K(\mathbf{w}_j(k))$ em relação a $\mathbf{w}_j$ resulta:

$$\frac{\partial K(\mathbf{w}_j(k))}{\partial \mathbf{w}_j} = -4\mathbf{R}_{\mathbf{x}_j}\mathbf{w}_j + 4(\mathbf{w}_j^T\mathbf{R}_{\mathbf{x}_r}\mathbf{w}_j)\mathbf{w}_j \tag{3.123}$$

Igualando a Equação 3.123 a zero, o mínimo de $K(\mathbf{w}_j(k))$ é dado por:

$$\mathbf{R}_{\mathbf{x}_j}\mathbf{w}_j \ = \lambda\mathbf{w}_j, \tag{3.124}$$

para:

$$\lambda = \mathbf{w}_j^T\mathbf{R}_{\mathbf{x}_r}\mathbf{w}_j \tag{3.125}$$

Se considerarmos $\mathbf{w}_j = \alpha\mathbf{e}_j$, ou seja, $\mathbf{w}_j$ igual ao j-ésimo autovalor normalizado de $\mathbf{R}_{\mathbf{x}}$, escalado pela constante arbitrária α, resulta para λ:

$$\lambda = \alpha(\mathbf{e}_j^T\mathbf{R}_{\mathbf{x}_r}\mathbf{e}_j^T)\alpha = \alpha^2\lambda_j, \tag{3.126}$$

e para a Equação 3.124:

$$\alpha\mathbf{R}_{\mathbf{x}_j}\mathbf{e}_j = \alpha^2\lambda_j\mathbf{e}_j \tag{3.127}$$

$$\mathbf{R}_{\mathbf{x}_j}\mathbf{e}_j = \alpha\lambda_j\mathbf{e}_j, \tag{3.128}$$

de onde se conclui que $\alpha = 1$, ou seja, os pesos ótimos de $K(\mathbf{w}_j(k))$ correspondem aos autovetores dominantes normalizados de $\mathbf{x}_j$. Dentre os $(N - j + 1)$ autovetores dominantes de $\mathbf{R}_{\mathbf{x}_j}$, o que minimiza a Equação 3.122 corresponde a:

$$\mathbf{w}_j = \mathbf{e}_j, \tag{3.129}$$

para o qual a Equação 3.122 resulta em:

$$K(\mathbf{w}_j(k)) = tr\{\mathbf{x}_s\mathbf{x}_s^T\} - \lambda_j, \tag{3.130}$$

ou seja, em relação aos pesos ótimos, é indiferente a escolha da modalidade de deflação .

3.7.1 Arquitetura de células auto-associativas com inibições laterais

Caso sejam estabelecidas, de forma apropriada, conexões laterais entre os neurônios da camada intermediária da hierarquia de células auto-associativas, é possível eliminar a estrutura de deflação apresentada na Figura 3.7.

Considere que a deflação é aplicada apenas às entradas. A saída do neurônio da j-ésima célula auto-associativa corresponde a:

$$y_j^{'} = \mathbf{w}_j^T \left(\mathbf{I} - \sum_{i=1}^{j-1} \mathbf{w}_i \mathbf{w}_i^T \right) \mathbf{x} \tag{3.131}$$

Desenvolvendo a equação anterior tem-se:

$$\begin{aligned} y_j^{'} &= \mathbf{w}_j^T \mathbf{x} - \mathbf{w}_j^T \sum_{i=1}^{j-1} \mathbf{w}_i y_i \\ y_j^{'} &= y_j - \sum_{i=1}^{j-1} c_{ji} y_i, \end{aligned} \tag{3.132}$$

para:

$$c_{ji} = \mathbf{w}_j^T \mathbf{w}_i \tag{3.133}$$

$$y_j = \mathbf{w}_j^T \mathbf{x} \tag{3.134}$$

Deste modo, uma forma computacionalmente econômica para a produção da deflação na entrada se resume nos seguintes passos:

1. Considerar o próprio vetor de dados como entrada de todas as células da hierarquia. Propagar este vetor por todas as células (Equação 3.134).

2. Produzir os valores de c_{ji} de cada célula (Equação 3.133).

3. Calcular o valor de $y_j^{'}$ por célula (Equação 3.132).

4. Utilizar na técnica de treinamento escolhida o valor de $y_j^{'}$, ao invês de y_j.

Esta forma é extremamente interessante quando o treinamento é por batelada, uma vez os valores de c_{ij} são calculados uma única vez a cada atualização de pesos, o que produz uma economia significativa de processamento na fase de treinamento. Na realidade, pela sistemática anterior, o que as constantes c_{ji} estão realizando é uma conexão entre as células auto-associativas, definindo uma nova estrutura que é ilustrada na Figura 3.9. É fácil verificar que a arquitetura proposta possui forte similaridade com a arquitetura de Rubner e Tavan [93] (vide seção 3.4.2.1).

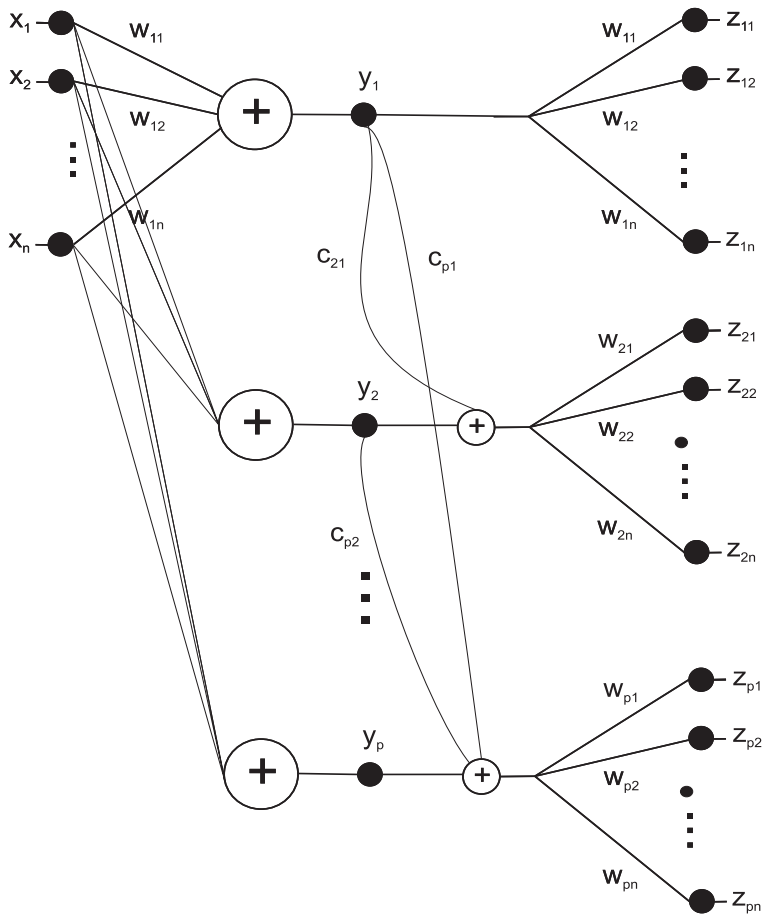


Figura 3.9: Hierarquia de células auto-associativas para a extração de componentes principais explorando a conexão lateral entre os neurônios

3.8 Modalidades de treinamento da hierarquia de células auto-associativas

Uma vez definida a arquitetura (conforme seção anterior), cabe a seleção de uma função objetivo e a aplicação de uma técnica de otimização para a extração das componentes principais. Com base em diferentes escolhas deste binômio, 6 novos algoritmos (3 de extração *on-line* e 3 de extração *off-line*) são propostos, os quais são apresentadas a seguir.

3.8.1 Métodos *on-line*

Dois são os representantes desta categoria: o método gradiente descendente estocástico e o método RLS.

3.8.1.1 Método gradiente descendente estocástico

Pela técnica de gradiente descendente [71], a atualização dos pesos da j-ésima célula auto-associativa deve ser realizada segundo a equação:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) - \eta(k)\nabla_{\mathbf{w}_j(k)}K(\mathbf{w}_j(k)), \tag{3.135}$$

onde $\eta(k)$ é uma constante arbitrária, referida como fator de aprendizado, que pode ser mantida fixa ou variar ao longo do treinamento, e $K(\mathbf{w}_j(k))$ corresponde a função objetivo dada pela Equação 3.118.

Pela Equação 3.118, $K(\mathbf{w}_j(k))$ consiste no valor esperado do erro quadrático. Para a produção de um algoritmo *on-line*, uma possibilidade é considerar, ao invés do valor esperado, o valor instantâneo do erro. Esta aproximação produz um algoritmo cuja atualização dos parâmetros é apenas dependente do último vetor de entrada, de forma similar aos filtros adaptativos LMS (*Least Mean Square*) [99]. Nesta aproximação, o gradiente utilizado para o treinamento é referido como estocástico [76]. Em resumo, para o cálculo do gradiente estocástico, aplica-se o gradiente sobre a função objetivo $K(\mathbf{w}_j(k))$ estimada segundo a forma:

$$K(\mathbf{w}_j(k)) \approx ||\mathbf{t}_j(k) - \mathbf{w}_j(k-1)\mathbf{w}_j^T(k-1)\mathbf{i}_j(k)||^2 \tag{3.136}$$

Neste caso, as equações de treinamento podem ser escritas como [86, 96]:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)[y_j(k)\mathbf{e}_j(k) + \mathbf{i}_j^T(k)\mathbf{e}_j(k)\mathbf{w}_j(k)], \tag{3.137}$$

para:

$$\mathbf{e}_j(k) = \mathbf{t}_j(k) - y_j(k)\mathbf{w}_j(k) \tag{3.138}$$

$$y_j(k) = \mathbf{w}_j^T(k)\mathbf{i}_j(k), \tag{3.139}$$

onde $\mathbf{i}_j(k)$ e $\mathbf{t}_j(k)$ são produzidos, de acordo com as Equações 3.112-3.117, utilizando $\mathbf{x}_j$ determinado pela Equação 3.111, de acordo com a modalidade de deflação selecionada.

Caso seja considerado que, na composição da Equação 3.137, o termo $\mathbf{i}_j^T(k)\mathbf{e}_j(k)\mathbf{w}_j(k)$ é pouco significativo em relação ao termo $y_j(k)\mathbf{e}_j(k)$, uma equação de treinamento alternativa é:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k)y_j(k)\mathbf{e}_j(k), \tag{3.140}$$

a qual aproxima a Equação 3.137. Esta aproximação é motivada pela verificação experimental, através de extensivas simulações, que o termo desprezado não intefere, significativamente, na acuidade e nem na velocidade de convergência, para as diferentes modalidades de deflação.

Há uma outra forma de derivar a Equação 3.140. Se a Equação 3.136 for reescrita na forma:

$$K(\mathbf{w}_j(k)) \approx ||\mathbf{t}_j(k) - y_j(k)\mathbf{w}_j(k)||^2, \tag{3.141}$$

onde:

$$y_j(k) = \mathbf{w}_j^T(k-1)\mathbf{i}_j(k), \tag{3.142}$$

está sendo considerado que o efeito sobre o valor de $y_j(k)$ da variação do peso entre as iterações $(k-1)$ e k não é significativo. Neste caso, o gradiente estocástico de $K(\mathbf{w}_j(k))$ assume a forma:

$$\nabla_{\mathbf{w}_j(k)} K(\mathbf{w}_j(k)) = -y_j(k)\mathbf{e}_j(k), \tag{3.143}$$

onde:

$$\mathbf{e}_j(k) = \mathbf{t}_j(k) - y_j(k)\mathbf{w}_j(k), \tag{3.144}$$

para o qual resultam as equações de treinamento na forma proposta pela Equação 3.140. Conclui-se que desprezar o termo $\mathbf{w}_j^T(k)\mathbf{e}_j(k)\mathbf{i}_j(k)$ na Equação 3.137 é equivalente a desprezar a variação das projeções entre iterações consecutivas.

Uma das vantagens desta modalidade de treinamento é permitir que a extração das componentes seja realizada de forma *on-line*. Em contrapartida, por ser uma técnica baseada em gradiente descendente estocástico, sua convergência é normalmente lenta e dependente da escolha de $\eta(k)$ [86], de forma similar a outros algoritmos da literatura baseados nesta técnica, conforme citado em [65]. A escolha de $\eta(k)$ influencia, também, na acuidade da extração. Em problemas práticos, para algoritmos da literatura baseados em gradiente estocástico, a escolha de $\eta(k)$ é realizada por tentativa e erro [100], fato crítico em aplicações de alta-dimensionalidade e grande número de eventos.

3.8.1.2 Método RLS

O objetivo da modalidade agora proposta é prover um algoritmo *on-line* que contorne os inconvenientes da modalidade baseada em gradiente estocástico, em especial, a velocidade de convergência e a necessidade da escolha do fator de aprendizado $\eta(k)$. Este método será baseado na otimização da seguinte função objetivo:

$$L(\mathbf{w}_j(k)) = \sum_{i=1}^{k} \beta^{k-i} |\mathbf{t}_j(i) - \mathbf{w}_j(k) y_j(k,i)|^2, \tag{3.145}$$

para:

$$y_j(k,i) = \mathbf{w}_j^T(k)\mathbf{i}_j(i), \tag{3.146}$$

onde β é uma constante, referida como fator de esquecimento, a ser escolhida no intervalo $0 < \beta \leq 1$. Neste caso, supõe-se que $L(\mathbf{w}_j(k))$ aproxima $K(\mathbf{w}_j(k))$ (Equação 3.118). Pode-se observar que esta equação é bastante similar à equação utilizada pelo método PAST (Equação 3.25), em especial, quando a última é considerada para a extração de uma única componente.

De forma análoga ao trabalho de Yang (1995) [69], para que seja possível determinar o valor ótimo de $\mathbf{w}_j(k)$ a cada passo de treinamento, será considerada a seguinte aproximação:

$$y_j(k,i) \approx y_j(i-1,i), \qquad 1 \leq i \leq k, \tag{3.147}$$

ou seja, que a variação das projeções dos dados no intervalo de iterações de 1 a k não foi significativa, resultando na seguinte função objetivo aproximada:

$$M(\mathbf{w}_j(k)) = \sum_{i=1}^{k} \beta^{k-i} |\mathbf{t}_j(i) - \mathbf{w}_j(k) y_j(i-1,i)|^2 \tag{3.148}$$

Expandindo a Equação 3.148, tem-se:

$$\begin{aligned} M(\mathbf{w}_j(k)) = & \sum_{i=1}^{k} \beta^{k-i} \mathbf{t}_j^T(i)\mathbf{t}_j(i) - 2\sum_{i=1}^{k} \beta^{k-i} y_j(i-1,i)\mathbf{t}_j^T(i)\mathbf{w}_j(k) \\ & + \sum_{i=1}^{k} \beta^{k-i} y_j^2(i-1,i)\mathbf{w}_j^T(k)\mathbf{w}_j(k) \end{aligned} \tag{3.149}$$

O gradiente de $M(\mathbf{w}_j(k))$ com relação a $\mathbf{w}_j(k)$ pode ser escrito como:

$$\nabla_{\mathbf{w}_j(k)} M(\mathbf{w}_j(k)) = -2\sum_{i=1}^{k} \beta^{k-i} \Big(\mathbf{t}_j(i) y_j(i-1,i) + y_j^2(i-1,i)\mathbf{w}_j(k) \Big) \tag{3.150}$$

Igualando a Equação 3.150 a zero, resulta na seguinte equação de treinamento:

$$\mathbf{w}_j(k) = \frac{\sum_{i=1}^{k} \beta^{k-i} \mathbf{t}_j(i) y_j(i-1,i)}{\sum_{i=1}^{k} \beta^{k-i} y_j^2(i-1,i)} \tag{3.151}$$

A forma apresentada pela Equação 3.151 não é conveniente para a implementação *on-line*, uma vez que o cálculo de $\mathbf{w}_j(k)$ envolve valores de y_j e $\mathbf{t}_j$ das iterações anteriores. Note porém que é possível reescrevê-la na forma:

$$\mathbf{w}_j(k) = \frac{\beta}{D_j(k)} \sum_{i=1}^{k-1} \beta^{(k-1)-i} y_j(i-1,i)\mathbf{t}_j(i) + \frac{1}{D_j(k)} y_j(k-1,i)\mathbf{t}_j(k) \tag{3.152}$$

para:

$$D_j(k) = \sum_{i=1}^{k} \beta^{k-i} y_j^2(i-1,i) \tag{3.153}$$

Pela Equação 3.151, a seguinte relação, derivada para a iteração $(k-1)$, é válida:

$$\sum_{i=1}^{k-1} \beta^{(k-1)-i} y_j(i-1,i)\mathbf{t}_j(i) = \mathbf{w}_j(k-1) \sum_{i=1}^{k-1} \beta^{(k-1)-i} y_j^2(i-1,i) \tag{3.154}$$

Substituindo a Equação 3.154 no primeiro termo da Equação 3.152, resulta:

$$\begin{aligned}
\mathbf{w}_j(k) &= \frac{\beta}{D_j(k)}\mathbf{w}_j(k-1)\sum_{i=1}^{k-1}\beta^{(k-1)-i}y_j^2(i-1,i) + \frac{1}{D_j(k)}y_j(k-1,k)\mathbf{t}_j(k) \\
\mathbf{w}_j(k) &= \frac{\mathbf{w}_j(k-1)}{D_j(k)}\left(\sum_{i=1}^{k}\beta^{k-i}y_j^2(i-1,i) - y_j^2(k-1,k)\right) + \\
&\quad \frac{1}{D_j(k)}y_j(k-1,k)\mathbf{t}_j(k) \\
\mathbf{w}_j(k) &= \frac{\mathbf{w}_j(k-1)}{D_j(k)}[D_j(k) - y_j^2(k-1,k)] + \frac{1}{D_j(k)}y_j(k-1,k)\mathbf{t}_j(k) \\
\mathbf{w}_j(k) &= \mathbf{w}_j(k-1) + \eta_j(k)y_j(k-1,k)\Big[\mathbf{t}_j(k) - y_j(k-1,k)\mathbf{w}_j(k-1)\Big],
\end{aligned} \tag{3.155}$$

onde $\eta_j(k) = \frac{1}{D_j(k)}$ [4]. Outra forma de escrever $\eta_j(k)$ é dada por:

$$\begin{aligned}
\frac{1}{\eta_j(k)} &= \beta\sum_{i=1}^{k-1}\beta^{(k-1)-i}y_j^2(i-1,i) + y_j^2(k-1,k) \\
\frac{1}{\eta_j(k)} &= \beta\frac{1}{\eta_j(k-1)} + y_j^2(k-1,k) \\
\eta_j(k) &= \frac{\eta_j(k-1)}{\beta + \eta_j(k-1)y_j^2(k-1,k)},
\end{aligned} \tag{3.156}$$

para a qual o valor de $\eta_j(k)$ pode ser determinado de forma recursiva, segundo uma fórmula de baixo custo computacional.

O interessante desta proposta é que a equação de treinamento obtida é idêntica a proposta pela Equação 3.140, sendo o fator de aprendizado determinado de forma ótima a cada iteração. Em razão da determinação adaptativa de $\eta_j(k)$, o algoritmo apresenta uma convergência significativamente mais rápida e acurada que o baseado em gradiente descendente estocástico. Faz-se necessário, no entanto, determinar um valor apropriado de β, o qual é dependente da aplicação.

Outro resultado interessante é que, como o valor $\eta_j(k)$ é dependente da iteração e da componente em extração, confirma-se, teoricamente, que uma escolha apropriada de $\eta_j(k)$ para a modalidade de gradiente descendente estocástico é também dependente destes fatores, conforme reportado pela literatura.

[4]Como é pressuposto que a matriz de correlação dos dados possui todos os autovalores distintos e não-nulos, $D_j(k) \neq 0$, visto que $D_j(k) \approx E[y_j^2]$ e $E[y_j^2] = \lambda_j$ (conforme Apêndice B), onde λ_j é o autovalor associado à j-ésima componente.

3.8.2 Métodos *off-line*

Para a derivação dos métodos desta categoria, o treinamento da j-ésima célula partirá da seguinte função objetivo:

$$\begin{aligned} L(\mathbf{w}_j(k)) &= \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{t}_j(i) - \mathbf{w}_j(k)\mathbf{w}_j^T(k)\mathbf{i}_j(i)||^2 \\ L(\mathbf{w}_j(k)) &= \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{t}_j(i) - y_j(k,i)\mathbf{w}_j(k)||^2, \end{aligned} \tag{3.157}$$

onde N é o número de eventos disponíveis e $y_j(k,i)$ é dado pela Equação 3.146. Trata-se, portanto, de uma aproximação da Equação 3.118, na qual o valor esperado do erro quadrático é estimado pelo seu valor médio para N eventos.

Uma aproximação conveniente [5] é considerar que a variação das projeções dos dados entre as iterações $(k-1)$ e k é desprezível, o que resulta:

$$M(\mathbf{w}_j(k)) = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{t}_j(i) - y_j(k-1,i)\mathbf{w}_j(k)||^2 \tag{3.158}$$

Desenvolvendo a equação anterior, tem-se:

$$\begin{aligned} M(\mathbf{w}_j(k)) = \frac{1}{N}\sum_{i=1}^{N}\Bigg(&\mathbf{t}_j^T(i)\mathbf{t}_j(i) - 2\sum_{i=1}^{N} y_j(k-1,i)\mathbf{t}_j^T(i)\mathbf{w}_j(k) \\ &+ \sum_{i=1}^{N} y_j^2(k-1,i)\mathbf{w}_j^T(k)\mathbf{w}_j(k)\Bigg), \end{aligned} \tag{3.159}$$

que pode ser reescrita na forma:

$$M(\mathbf{w}_j(k)) = \mathbf{c}(k) - 2\mathbf{b}^T(k)\mathbf{w}(k) + \mathbf{w}^T(k)\mathbf{A}\mathbf{w}(k) \tag{3.160}$$

onde:

$$\mathbf{A}(k) = \frac{1}{N}\sum_{i=1}^{N} y_j^2(k-1,i)\mathbf{I} \tag{3.161}$$

$$\mathbf{b}(k) = \frac{1}{N}\sum_{i=1}^{N} y_j(k-1,i)\mathbf{t}_j(k) \tag{3.162}$$

$$\mathbf{c}(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{t}_j^T(i)\mathbf{t}_j(i), \tag{3.163}$$

[5] Cabe observar que esta aproximação foi discutida na derivação dos algoritmos *on-line* baseados em gradiente estocástico.

onde $\mathbf{I}$ é a matriz identidade. Note ainda que minimizar a Equação 3.160 é equivalente a minimizar a seguinte equação:

$$M(\mathbf{w}_j(k)) \quad = \tfrac{1}{2}\mathbf{w}^T(k)\mathbf{A}\mathbf{w}(k) - \mathbf{b}^T(k)\mathbf{w}(k), \tag{3.164}$$

que difere da primeira forma por uma normalização e constante arbitrária. O gradiente - $\nabla_{\mathbf{w}_j(k)}M(\mathbf{w}_j(k))$ - e a hessiana - $H_{M(\mathbf{w}_j(k))}$ - da Equação 3.164 são dados por:

$$\nabla_{\mathbf{w}_j(k)}M(\mathbf{w}_j(k)) = \mathbf{A}\mathbf{w} - \mathbf{b} = -\frac{1}{N}\sum_{i=1}^{N} y_j(k-1,i)[\mathbf{t}_j(k) - \mathbf{w}_j(k)y_j(k-1,i)] \tag{3.165}$$

$$H_{M(\mathbf{w}_j(k))} = \mathbf{A} = \frac{1}{N}\sum_{i=1}^{N} y_j^2(k-1,i)\mathbf{I} \tag{3.166}$$

Aspecto interessante é que a matriz Hessiana é diagonal, determinada pelo valor médio do quadrado das projeções. Assim, utilizando a função objetivo da Equação 3.164 e os valores do seu gradiente e hessiana, serão derivados os métodos a seguir.

3.8.2.1 Método gradiente descendente com passo ótimo

Segundo a técnica de gradiente descendente [71], a atualização dos pesos da j-ésima célula auto-associativa pode ser descrita pela equação iterativa:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \alpha_j(k)\mathbf{d}_j(k), \tag{3.167}$$

onde $\mathbf{d}_j(k)$ constitui a direção de procura do algoritmo, a qual é escolhida como:

$$\mathbf{d}_j(k) = -\nabla_{\mathbf{w}_j(k)}M(\mathbf{w}_j(k)), \tag{3.168}$$

ou seja, $\mathbf{d}_j$ é escolhido como a direção de maior decréscimo de $M(\mathbf{w}_j(k))$.

Uma escolha interessante para $\alpha_j(k)$ é aquela que minimiza o valor de $M(\mathbf{w}_j(k+1))$. Para esta escolha de $\alpha_j(k)$, o algoritmo realiza um passo ótimo, devendo a seguinte equação ser satisfeita:

$$\nabla_{\alpha(k)}M(\mathbf{w}_j(k+1)) = 0 \tag{3.169}$$

Segundo a Equação 3.164, o valor de $M(\mathbf{w}_j(k+1))$ pode ser expresso na forma [6]:

$$\begin{aligned} M(\mathbf{w}_j(k+1)) &= \frac{1}{2}[\mathbf{w}_j + \alpha_j \mathbf{d}_j]^T \mathbf{A}[\mathbf{w}_j + \alpha_j \mathbf{d}_j] - \mathbf{b}^T[\mathbf{w}_j + \alpha_j \mathbf{d}_j] \\ M(\mathbf{w}_j(k+1)) &= \frac{1}{2}\mathbf{w}_j^T \mathbf{A}\mathbf{w}_j - \mathbf{b}^T\mathbf{w}_j + \frac{1}{2}\alpha_j^2 \mathbf{d}_j^T \mathbf{A}\mathbf{d}_j + \alpha_j \mathbf{d}_j^T \mathbf{A}\mathbf{w}_j - \alpha_j \mathbf{b}^T \mathbf{d}_j \end{aligned} \tag{3.170}$$

Substituindo a Equação 3.170 na 3.169, resulta:

$$\begin{aligned} \mathbf{d}_j^T \mathbf{A}\mathbf{w}_j + \alpha_j \mathbf{d}_j^T \mathbf{A}\mathbf{d}_j - \mathbf{b}^T \mathbf{d}_j = 0 \\ \alpha_j = \frac{\mathbf{d}_j^T(\mathbf{b} - \mathbf{A}\mathbf{w}_j)}{\mathbf{d}_j^T \mathbf{A}\mathbf{d}_j} = -\frac{\mathbf{d}_j^T \mathbf{d}_j}{\mathbf{d}_j^T \mathbf{A}\mathbf{d}_j} \end{aligned} \tag{3.171}$$

Substituindo a Equação 3.165 e 3.166 na 3.171, resulta que o valor ótimo de $\alpha_j(k)$ é dado por:

$$\alpha_j(k) = -\frac{1}{\frac{1}{N}\sum_{i=1}^{N} y_j^2(k-1,i)}, \tag{3.172}$$

e a equação de treinamento assume a forma:

$$\begin{aligned} \mathbf{w}_j(k) = \mathbf{w}_j(k-1) \\ - \alpha(k)\frac{1}{N}\sum_{i=1}^{N} y_j(k-1,i)[\mathbf{t}_j(k) - \mathbf{w}_j(k-1)y_j(k-1,i)] \end{aligned} \tag{3.173}$$

Um outro caminho para a derivação deste algoritmo consiste em calcular, analiticamente, o valor de $\mathbf{w}_j(k)$ para o qual $M(\mathbf{w}_j(k+1))$ é mínimo, ou seja, que satisfaz: $\nabla_{\mathbf{w}_j(k)} M(\mathbf{w}_j(k+1)) = 0$. Pela Equação 3.165, o valor ótimo de $\mathbf{w}_j(k)$ é dado por: $\mathbf{w}_j(k) = \mathbf{A}^{-1}(k)\mathbf{b}(k)$, o que resulta na seguinte equação de treinamento:

$$\mathbf{w}_j(k) = \frac{\sum_{i=1}^{N} y_j(k-1,i)\mathbf{t}_j(i)}{\sum_{i=1}^{N} y_j^2(k-1,i)} \tag{3.174}$$

Note que é possível reescrever a Equação 3.174 na forma:

$$\begin{aligned} \mathbf{w}_j(k) = \frac{1}{D_j(k)}\left(D_j(k)\mathbf{w}_j(k-1) + \frac{1}{N}\sum_{i=1}^{N} y_j(k-1,i)\mathbf{t}_j(i) - D_j(k)\mathbf{w}_j(k-1) \right) \\ \mathbf{w}_j(k) = \mathbf{w}_j(k-1) + \frac{1}{D_j(k)}\frac{1}{N}\sum_{i=1}^{N} y_j(k-1,i)[\mathbf{t}_j(k) - \mathbf{w}_j(k-1)y_j(k-1,i)], \end{aligned} \tag{3.175}$$

para

$$D_j(k) = \frac{1}{N}\sum_{i=1}^{N} y_j^2(k-1,i), \tag{3.176}$$

donde concluímos que ambas derivações produzem as mesmas equações de treinamento. Um dos principais atrativos do algoritmo proposto é que nenhum parâmetro ou fator de aprendizado é necessário para a extração.

[6] Para facilitar a leitura foram omitidos de $\mathbf{w}_j$, α_j e d_j a iteração (k) correspondente.

3.8.2.2 Gradiente conjugado

A técnica de gradiente conjugado é tida, normalmente, como um método intermediário entre a técnica de gradiente descendente com passo ótimo e o método de Newton, em termos de complexidade e velocidade de convergência [101]. Sua principal vantagem é que a matriz hessiana não precisa ser avaliada, armazenada e invertida como no método de Newton. Trata-se de uma classe de algoritmos hábeis em lidar com problemas de larga escala de forma eficiente [102].

Desenvolvido originalmento por Hestenes e Stiefel (1952) [103], para a solução de um conjunto de equações com uma matriz de coeficientes positiva e definida, sua característica fundamental é o fato que as direções percorridas pelo algoritmo ao longo do processo de otimização são ortogonais entre si com respeito a uma matriz arbitrária. Este fato resulta que a otimização realizada numa dada iteração não "estraga"as otimizações das iterações anteriores, de forma diferente à técnica de *Steepest descent*. Pode-se mostrar ainda que para problemas quadráticos, a técnica converge em N passos, onde N corresponde ao número de parâmetros envolvidos [104].

A aplicação da técnica de gradiente conjugado à extração de autovetores já é conhecida da comunidade de processamento de sinais. Entre trabalhos relevantes, tem-se Yang, Sakar e Arvas (1989) [105] e Fu e Dowling (1995) [106], entre outros. Ambos trabalhos exploram a minimização do coeficiente de Rayleigh, o último aplicando a deflação para extração de múltiplas componentes. Nossa proposta se diferencia das anteriores por realizar a extração através da hierarquia de células auto-associativas, considerando a otimização da função objetivo da Equação 3.164.

Considere uma função objetivo arbitrária: $J_j(\mathbf{w}_j(k))$. A técnica de gradiente conjugado, quando aplicada a sua otimização, resulta nas seguintes equações [107]:

$$\mathbf{d}_j(0) = \mathbf{r}_j(0) = -\nabla_{\mathbf{w}_j(k)} J_j(\mathbf{w}_j(0)) \tag{3.177}$$

$$\alpha_j(k) = \frac{-\mathbf{r}_j^T(k-1)\mathbf{d}_j(k-1)}{\mathbf{d}_j^T(k-1)\mathbf{H}_{J_j(\mathbf{w}_j(k))}\mathbf{d}_j(k-1)} \tag{3.178}$$

$$\mathbf{w}_j(k) = \mathbf{w}_j(k-1) + \alpha_j(k)\mathbf{d}_j(k-1) \tag{3.179}$$

$$\mathbf{r}_j(k) = -\nabla_{\mathbf{w}_j(k)} J_j(\mathbf{w}_j(k+1)) \tag{3.180}$$

$$\beta_j(k) = \frac{\mathbf{r}_j^T(k)\mathbf{r}_j(k)}{\mathbf{r}_j^T(k-1)\mathbf{r}_j(k-1)} \tag{3.181}$$

$$\mathbf{d}_j(k) = \mathbf{r}_j(k) + \beta_j(k)\mathbf{d}_j(k-1), \tag{3.182}$$

onde $\nabla_{\mathbf{w}_j(k)} J_j(\mathbf{w}_j(k))$ e $\mathbf{H}_{J_j(\mathbf{w}_j(k))}$ são, respectivamente, o gradiente e a hessiana da função $J_j(\mathbf{w}_j(k))$.

A Equação 3.181 foi proposta por Fletcher e Reeves (1964) [107]. Outra possibilidade deve-se a Polak e Ribière (1971) [108], sendo dada por:

$$\beta_j(k) = \frac{\mathbf{r}_j^T(k)(\mathbf{r}_j(k) - \mathbf{r}_j(k-1))}{\mathbf{r}_j^T(k-1)\mathbf{r}_j(k-1)}, \tag{3.183}$$

que apresenta uma maior velocidade de convergência [104].

Para a função objetivo da Equação 3.164, substituindo os valores de gradiente e hessiana dados pelas Equações 3.165 e 3.166 no conjunto das Equações 3.177-3.182, resulta que as equações de treinamento do método proposto são dadas por:

$$\mathbf{d}_j(0) = \mathbf{r}_j(0) = \frac{1}{N}\sum_{i=1}^{N} y_j(0,i)[\mathbf{t}_j(i) - y_j(0,i)\mathbf{w}_j(0)] \tag{3.184}$$

$$\alpha_j(k) = \frac{-\mathbf{r}_j^T(k-1)\mathbf{d}_j(k-1)}{|\mathbf{d}_j(k-1)|^2 \sum_{i=1}^{N} y_j^2(k-1,i)} \tag{3.185}$$

$$\mathbf{w}_j(k) = \mathbf{w}_j(k-1) + \alpha_j(k)\mathbf{d}_j(k-1) \tag{3.186}$$

$$\mathbf{r}_j(k) = \frac{1}{N}\sum_{i=1}^{N} y_j(k,i)[\mathbf{t}_j(k) - y_j(k,i)\mathbf{w}_j(k)] \tag{3.187}$$

$$\mathbf{d}_j(k) = \mathbf{r}_j(k) + \beta_j(k)\mathbf{d}_j(k-1), \tag{3.188}$$

para $\beta_j(k)$ dado pela Equação 3.181 ou 3.183, e $y_j(k,i)$ dado pela Equação 3.146.

Um procedimento recomendável para o método de gradiente conjugado é reinicializar sua direção de procura periodicamente, uma vez que por erros numéricos há uma degradação de ortogonalidade entre as direções de procura. Um procedimento interessante deve-se a Powell (1977) [109], que reinicializa esta direção para o negativo do gradiente quando o valor de $\beta_j(k)$ (Equação 3.181 ou 3.183) é negativo.

3.8.2.3 Método baseado no treinamento iRPROP

Uma técnica interessante para o treinamento da hierarquia de células auto-associativas é o *backpropagation* resiliente, proposto por Riedmiler em duas versões, uma delas em 1993 [110]; e outra, em 1994 [111]. Nesta técnica, evitam-se eventuais paralisias do treinamento em regiões acentuadamente planas da função objetivo, uma vez que o processo de treinamento é apenas dependente do sinal de sua derivada. Este método é comumente referido como o melhor método de primeira ordem para

o treinamento de redes neurais. Trata-se de uma técnica que conjuga simplicidade, baixo custo computacional por iteração e eficácia, e seu desempenho é comparável a algoritmos mais complexos e de custo computacional significativamente maior, tais como o BFGS [112]. Várias variantes da técnica foram ainda propostas, destacando-se o iRPROP, de autoria de Igel e Husken (2000) [113].

De forma análoga a derivação por gradiente conjugado, nossa proposta será aplicar o algoritmo iRPROP para o treinamento da hierarquia de células auto-associativas [114], utilizando a função objetivo da Equação 3.158, cujo gradiente, conforme discussão anterior, é dado pela Equação 3.165.

O algoritmo iRPROP realiza a minimização de cada parâmetro da função objetivo de forma independente. Sejam $\Delta_{ij}(k)$ e $\Delta\mathbf{w}_j(k)_i$ o valor do passo e o incremento, respectivamente, da i-ésima componente do j-ésimo vetor de pesos da hierarquia. A atualização dos parâmetros é realizada de acordo com a evolução das derivadas entre duas iterações consecutivas. Três são as situações possíveis: derivadas com sinais iguais, derivadas com sinais opostos, e pelo menos uma derivada nula. Quando os sinais são iguais, há um indicativo que o algoritmo caminha em direção ao mínimo, logo a componente do peso $\mathbf{w}_j(k)_i$ é atualizada, e o valor do passo corrente a ela associado ($\Delta_{ij}(k)$) é aumentando. Quando os sinais são diferentes, tem-se um indicativo que o valor do último passo foi excessivo, e o algoritmo distanciou-se de seu mínimo local. Neste caso, reduz-se o valor do passo e o deslocamento pode ou não ser desfeito, de acordo com a evolução do erro entre as duas últimas iterações. Havendo um crescimento do erro, o deslocamento é desfeito, em caso contrário, mantido. Por fim, quando pelo menos uma das derivadas é nula, mantém-se o valor do passo da iteração anterior. Este algoritmo pode ser resumido nos seguintes passos:

Para cada $\left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)}\right)_i$:

se: $\left(\frac{\partial M(\mathbf{w}_j(k-1))}{\partial \mathbf{w}_j(k-1)}\right)_i \left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)}\right)_i > 0$ **faça:**

$$\Delta_{ij}(k) = \text{mín}(\Delta_{ij}(k-1)\eta_+, \Delta_{\text{máx}})$$

$$\Delta \mathbf{w}_j(k)_i = -\text{sign}\left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)_i}\right)\Delta_{ij}(k)$$

$$\mathbf{w}_j(k+1)_i = \mathbf{w}_j(k)_i + \Delta \mathbf{w}_j(k)_i$$

caso contrário, se: $\left(\frac{\partial M(\mathbf{w}_j(k-1))}{\partial \mathbf{w}_j(k-1)}\right)_i \left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)}\right)_i < 0$ **faça:**

$$\Delta_{ij}(k) = \text{máx}(\Delta_{ij}(k-1)\eta_-, \Delta_{\text{mín}}) \tag{3.189}$$

se $M(\mathbf{w}_j(k)) > M(\mathbf{w}_j(k-1))$ **faça** $\mathbf{w}_j(k+1)_i = \mathbf{w}_j(k)_i - \Delta \mathbf{w}_j(k-1)_i$

$$\left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)}\right)_i = 0$$

caso contrário, se: $\left(\frac{\partial M(\mathbf{w}_j(k-1))}{\partial \mathbf{w}_j(k-1)}\right)_i \left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)}\right)_i = 0$ **faça:**

$$\Delta \mathbf{w}_j(k)_i = -\text{sign}\left(\frac{\partial M(\mathbf{w}_j(k))}{\partial \mathbf{w}_j(k)}\right)_i \Delta_{ij}(k)$$

$$\mathbf{w}_j(k+1)_i = \mathbf{w}_j(k)_i + \Delta \mathbf{w}_j(k)_i$$

final do se,

onde $M(\mathbf{w}_j(k))$ é a função objetivo associada ao treinamento da j-ésima célula auto-associativa, $sign(x)$ é a função sinal [7] e η_+, η_-, $\Delta_{\text{mín}}$ e $\Delta_{\text{máx}}$ são constantes do algoritmo. Em [110], os seguintes valores são sugeridos: $\eta_+ = 1,2$, $\eta_- = 0,5$, $\Delta_{\text{mín}} = 10^{-6}$ e $\Delta_{\text{máx}} = 50$. Como valor inicial dos incrementos, costuma-se utilizar $\Delta_0 = 0.05$. Para grande parte das aplicações, a técnica mostra-se ainda pouco sensível a escolha destes parâmetros [110, 111].

3.8.2.4 Método baseado na otimização por Newton-Rapson

O método de Newton-Rapson, aplicado a otimização de uma função objetivo arbitrária $J(\mathbf{w}_j(k))$, é baseado na seguinte equação iterativa:

$$\mathbf{w}_j(k) = \mathbf{w}_j(k-1) - (H_{J(\mathbf{w}_j(k))})^{-1} \nabla_{\mathbf{w}_j(k)} J(\mathbf{w}_j(k)), \tag{3.190}$$

[7] A função sinal é definida como: se $x > 0$, $sign(x) = 1$; se $x < 0$, $sign(x) = -1$.

onde $\nabla_{\mathbf{w}_j(k)} J(\mathbf{w}_j(k))$ e $H_{J(\mathbf{w}_j(k))}$ são, respectivamente, o gradiente e a hessiana de $J(\mathbf{w}_j(k))$.

Considerando a função objetivo, o gradiente e a hessiana dados pelas Equações 3.158, 3.165 e 3.166, respectivamente, as equações de treinamento para a hieraquia de células auto-associativas para o método de Newton resultam em:

$$\begin{aligned}\mathbf{w}_j(k) = \mathbf{w}_j(k-1) + \\ \frac{1}{D_j(k)} \frac{1}{N} \sum_{i=1}^{N} y_j(k-1,i)[\mathbf{t}_j(i) - \mathbf{w}_j(k) y_j(k-1,i)]\end{aligned} \tag{3.191}$$

para

$$D_j(k) = \sum_{i=1}^{N} y_j(k-1,i)^2 \tag{3.192}$$

Comparando as Equações 3.191 e 3.173, verifica-se que são idênticas, logo as derivações por gradiente descendente com passo ótimo e Newton-Rapson são equivalentes, o que é uma conseqüência da superfície de erro aproximada ser quadrática.

Na Tabela 3.4 resumimos os métodos propostos, apresentando seu domínio de aplicação (*on-line* x *off-line*), a função objetivo considerada em sua derivação e as equações de treinamento produzidas.

Tabela 3.4: Algoritmos propostos para a extração de componentes principais. Veja o texto.

Nome	**Aplicação**	**Função Objetivo**	**Equações**
Gradiente descendente estocástico	*on-line*	3.136	3.137-3.139
Gradiente descendente estocástico aproximado	*on-line*	3.141	3.140
Gradiente descendente passo ótimo	*off-line*	3.158	3.173
Gradiente conjugado	*off-line*	3.158	3.184-3.188
RPROP	*off-line*	3.158	3.189
Newton-Rapson	*off-line*	3.158	3.191- 3.192

3.9 Similaridades com métodos da literatura

Conforme mencionado anteriormente, algumas das modalidades propostas apresentam equações de treinamento similares a métodos relevantes da literatura. Estas relações serão exploradas a seguir.

3.9.1 *Generalized Hebbian Algorithm*

Se considerarmos a modalidade gradiente estocástico, na forma da Equação 3.140, caso a deflação seja aplicada apenas aos vetores-alvo, ou seja: $\mathbf{i}_j(k) = \mathbf{x}(k)$ e $\mathbf{t}_j(k) = \mathbf{x}_j(k)$, resulta:

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \eta(k) y_j(k) [\mathbf{x}_j(k) - y_j(k)\mathbf{w}_j(k)] \tag{3.193}$$

$$y_j(k) = \mathbf{w}_j^T(k)\mathbf{x}(k) \tag{3.194}$$

$$\mathbf{x}_j(k) = \mathbf{x}(k) - \sum_{i=1}^{j-1} y_i(k)\mathbf{w}_j(k) \tag{3.195}$$

Comparando as Equações 3.193-3.195 com as Equações 3.68-3.70, verificamos que elas são idênticas. Desta forma, o algoritmo GHA [38] pode ser derivado através do treinamento da hierarquia de células auto-associativas por gradiente descendente estocástico, considerando a aproximação discutida na seção 3.8.1.1.

3.9.2 Redes RLS

Se tomarmos a Equação 3.155 e considerarmos a deflação aplicada apenas aos vetores-alvo, ou seja, $\mathbf{i}_j(k) = \mathbf{x}(k)$ e $\mathbf{t}_j(k) = \mathbf{x}_j(k)$, assim como o valor de $\beta = 1$, resulta:

$$\mathbf{w}_j(k) = \mathbf{w}_j(k-1) + \eta(k) y_j(k-1,k) \Big[\mathbf{x}_j(k) - y_j(k-1,k)\mathbf{w}_j(k-1)\Big] \tag{3.196}$$

$$y_j(k-1,k) = \mathbf{w}_j^T(k-1)\mathbf{x}(k) \tag{3.197}$$

$$\mathbf{x}_j(k) = \mathbf{x}(k) - \sum_{i=1}^{j-1} y_i(k-1,k)\mathbf{w}_j(k-1) \tag{3.198}$$

$$\eta_j(k) = \frac{\eta_j(k-1)}{1 + \eta_j(k-1) y_j^2(k-1,k)} \tag{3.199}$$

Inicialmente vamos considerar o método de Bannour e Azimi-Sadjadi (1995) [82]. Retomando a Equação 3.85, com base na Equação 3.83, tem-se:

$$\begin{aligned} P_j(n) &= \left[1 - \frac{P_j(n-1)h_j^2(n)}{1 + P_j(n-1)h_j^2(n)}\right] P_j(n-1) \\ P_j(n) &= \frac{P_j(n-1)}{1 + h_j^2(n)P_j(n-1)} \end{aligned} \tag{3.200}$$

Note ainda que a Equação 3.83, com base na Equação 3.200, pode ser escrita na forma:

$$K_j(n) = P_j(n)h_j(n) \tag{3.201}$$

Assim, o método de Bannour [82] pode ser reescrito como:

$$h_j(n) = \mathbf{w}_j^T(n-1)\mathbf{x}(n) \tag{3.202}$$

$$P_j(n) = \frac{P_j(n-1)}{[1 + h_j^2(n)P_j(n-1)]} \tag{3.203}$$

$$\mathbf{w}_j(n) = \mathbf{w}_j(n-1) + P_j(n)h_j(n)[\mathbf{d}_j(n) - h_j(n)\mathbf{w}_j(n-1)], \tag{3.204}$$

para $\mathbf{d}_j(n)$ dado pela Equação 3.23.

Comparando as Equações 3.202-3.204 com as Equações 3.196-3.199, verifica-se que são idênticas, donde se conclui que a técnica de Bannour [82] também pode ser derivada considerando o treinamento da hierarquia de células auto-associativas, segundo a modalidade RLS, quando a deflação é aplicada apenas aos vetores-alvo e valor de β vale um. Relação similar pode ser estabelecida com a técnica PASTd, considerando a deflação aplicada, simultaneamente, aos vetores-alvo e às entradas.

3.9.3 Método das potências

Se considerarmos o treinamento das células auto-associativas pelo gradiente descendente com passo ótimo, para uma deflação aplicada, simultaneamente, às entradas e aos vetores alvos, resulta numa equação de treinamento dada por:

$$\mathbf{w}_j(k) = \frac{\sum_{i=1}^{N} y_j(k-1,i)\mathbf{x}_j(i)}{\sum_{i=1}^{N} y_j^2(k-1,i)} \tag{3.205}$$

Comparando a Equação 3.205 com a 3.106, verificamos que são bastante similares, exceto pela normalização, que é realizada a cada passo. Vale observar que a normalização imposta ao método das potências foi arbitrária, o que aproxima os

dois algoritmos. A vantagem da Equação 3.174 é um custo computacional inferior, em especial por não exigir o cálculo da raiz quadrada envolvida no módulo, fato atraente para implementações em microcontroladores ou processadores digitais de sinais.

3.10 Resultados comparativos de acuidade e custo computacional entre os algoritmos propostos e técnicas da literatura

Nesta análise buscou-se avaliar a aplicação dos métodos propostos à extração de características a serem utilizadas pelo sistema de classificação automática dos contatos. Estas características podem ser empregadas tanto na fase de operação do sistema, através de componentes extraídas de forma *off-line*, quanto no processo de adaptação/inclusão de novas classes, onde uma extração *on-line* das componentes pode ser necessária. A análise dos algoritmos com base em dados reais, que apresentam alta-dimensionalidade, larga flutuação estatística e grande número de eventos é também um interessante teste de robustez, que pode orientar a seleção de uma técnica específica para outros problemas e conjuntos de dados.

Os métodos propostos foram avaliados, comparativamente, com métodos da literatura em dois aspectos: a acuidade de extração e a velocidade de convergência, o último estimado pelo número de iterações necessárias à extração, utilizando uma partição dos dados do sonar representativa ao problema, composta por 7597 eventos de 557 componentes [8]. Para a análise de acuidade dos métodos foi realzada uma extração de referência, com a qual a direção das componentes fornecidas por cada método foi comparada. Esta extração de referência foi baseada numa implementação do pacote *eispack* [115], amplamente utilizado pela comunidade científica [9], em

[8] Esta partição foi produzida por sorteio, conforme processo descrito na seção 4.3, e considerou uma seleção por espectros (vide seção 2.1.3).

[9] O pacote *eispack* destina-se a uma extração *off-line* de todas as componentes do processo aleatório de interesse, explorando técnicas da álgebra linear numérica, entre elas: rotações e fatorações que operam sobre a matriz de correlação do processo visando diagonalizá-la. Para maiores detalhes, consultar [97]

razão de sua precisão e acuidade. Os autovalores de referência e a curva de energia associada são apresentadas na Figura 3.10 e na Tabela 3.5. Através da curva de energia, é possível identificar o percentual de energia do processo original que é retido em dados compactados segundo dado número de componentes. Por esta curva, são necessárias em torno de 300 componentes para reter um valor superior a 90% da energia do processo. A razão entre o maior e menor autovalor (condicionamento da matriz de correlação) é de 2372,5 , enquanto o menor autovalor vale: $\approx 1,7E-5$.

Com relação à análise de acuidade, são comuns na literatura medidas baseadas no erro de reconstrução dos dados com base nas componentes extraídas, aferido através do erro médio quadrático ou da relação sinal-ruído, e na comparação das componentes extraídas com as de referência, através da diferença quadrática, ângulo ou valor do autovalores estimados. A análise realizada com os métodos propostos enfocou três aspectos:

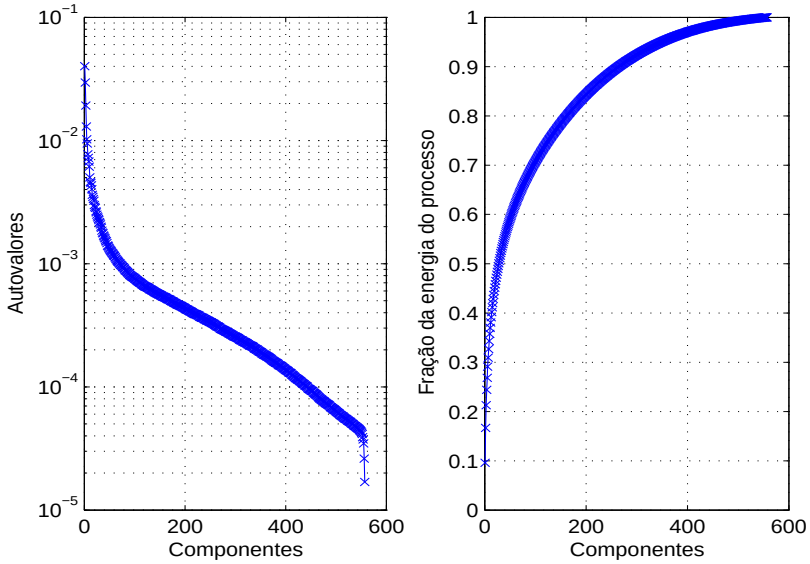


Figura 3.10: Autovalores e curva de energia da extração de referência.

- **Acuidade das direções das componentes extraídas:** foi avaliado o ângulo (em graus) entre cada componente extraída e a componente de referência correspondente.

Tabela 3.5: Alguns autovalores e valores de energia para a extração de referência.

Componente	**Autovalor**	**Energia acumulada (%)**	**Componente**	**Autovalor**	**Energia acumulada (%)**
1	4,0E-2	9,6	**300**	2,6E-4	92,4
50	1,3E-3	59,0	**400**	1,4E-4	97,0
100	7,6E-4	70,7	**500**	6,4E-5	99,3
200	4,3E-4	84,3	**557**	1,7E-5	100,0

- **Ortogonalidade de cada componente em relação às demais:** para a estimativa da j-ésima componente foi considerada a seguinte medida:

$$\phi_j = 90 - arccos\left(\sqrt{\frac{1}{556}\sum_{i\neq j, j=1}^{556}\theta_{ij}^2}\right) \tag{3.206}$$

para

$$\theta_{ij} = \frac{\mathbf{w}_i^T \mathbf{w}_j}{|\mathbf{w}_i||\mathbf{w}_j|}, \tag{3.207}$$

onde $arccos$ é a função arco-cosseno. A medida ϕ_j corresponde a um desvio médio absoluto de ortogonalidade entre a j-ésima componente e as demais componentes extraídas. Para a j-ésima componente, este valor é calculado com base na raiz do valor médio do quadrado dos desvios verificados entre os 556 pares de componentes (j, i) disponíveis. Conclui-se que quanto pior for a ortogonalidade das componentes extraídas, maior será o valor de ϕ_j. Caso as componentes sejam perfeitamente ortogonais, $\phi_j = 0$.

- **Acuidade da estimação dos autovalores com base nas componentes:** para cada componente extraída ($\mathbf{w}_i$), o autovalor associado foi estimado pela fórmula:

$$\tilde{\lambda}_i = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbf{w}_i^T}{|\mathbf{w}_i|}\mathbf{x}_i\right)^2, \tag{3.208}$$

onde N é o número de eventos ($\mathbf{x}_i$) disponível (vide Apêndice B). Esta análise considerou o erro relativo entre os autovalores estimados e os de referência.

Com relação à implementação, cada método foi codificado em linguagem C ANSI. A extração realizada foi seqüêncial, isto é, o método só inicia a extração da

segunda componente após a conclusão da extração da primeira, e assim, sucessivamente. Uma componente é considerada extraída quando a rede satisfaz um dos dois critérios de parada utilizados: a estabilização do autovalor da componente em extração e o número máximo de passos. Pelo primeiro, a cada iteração, determina-se a variação do autovalor estimado com respeito ao autovalor estimado na iteração anterior. Caso esta variação seja inferior a um limite especificado, incrementa-se um contador. Completado um número arbitrado de falhas consecutivas, a componente é considerada extraída. A utilização do contador de falhas consecutivas evita que o processo de extração seja interrompido de forma prematura em momentos críticos do processo de otimização, como, por exemplo, em regiões demasiado planas da função objetivo. Através de um pequeno número de simulações foi verificado que um valor apropriado para o limite de variação dos autovalores é de 10^{-14}, que está relacionado à acuidade máxima do ambiente de implementação. Para o limite de falhas, o valor escolhido foi de 10, enquanto para o número máximo de passos foi arbitrado o valor de 10000. Um dos atrativos do critério de convergência baseado na variação de autovalor é sua simplicidade e baixo custo computacional, em especial quando comparado com outros critérios como, por exemplo, a variação dos pesos entre iterações consecutivas.

Conforme a discussão da seção 3.6, para a análise de desempenho foram selecionados os seguintes métodos da literatura: GHA, PASTd e das potências. Em relação aos algoritmos propostos, a análise considerou os métodos gradiente conjugado, RPROP e Newton [10].

As análises foram divididas em três etapas: uma primeira envolvendo apenas métodos *off-line*, realizada entre os métodos gradiente conjugado, RPROP, Newton e o método das potências; uma segunda, onde o método de melhor desempenho da primeira análise é comparado com os métodos GHA e PASTd; e, por fim, a terceira, onde o método proposto de melhor desempenho é avaliado com respeito a diferentes

[10]Cabe observar, conforme a seção 3.8.2.4, que a derivação das equações de treinamento da hierarquia de células auto-associativas por gradiente descendente com passo-ótimo produz equações idênticas a derivação pelo método de Newton-Rapson, o que justifica a não existência de simulações contemplando o primeiro caso. De forma similar, não foram incluídas simulações com os métodos propostos gradiente descendente estocástico e RLS, pela sua similaridade com os métodos GHA e PASTd, respectivamente (vide seção 3.9)

modalidades de deflação. Para permitir uma justa comparação entre a velocidade de convergência dos diferentes algoritmos, todos os métodos consideraram os mesmos valores para os vetores iniciais das componentes.

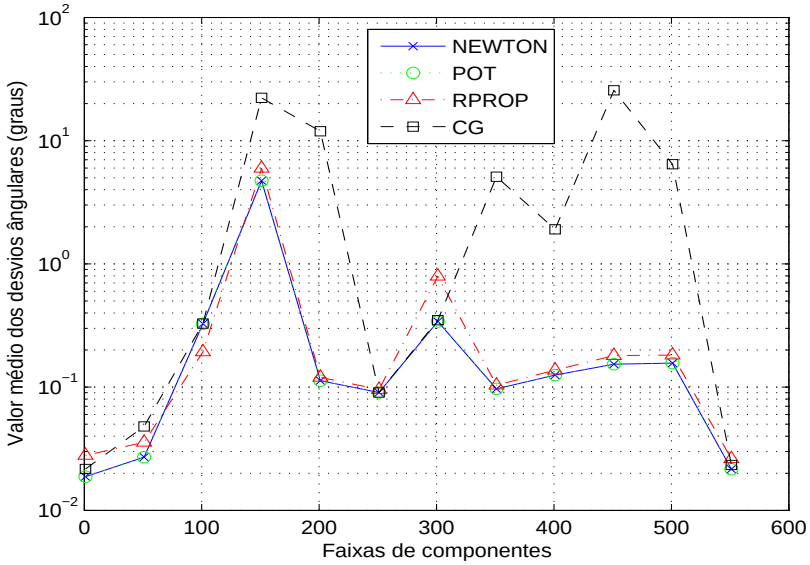
3.10.1 Comparação entre os métodos gradiente conjugado, RPROP e Newton com o método das potências

Nesta comparação, os métodos propostos utilizaram uma deflação aplicada, simultaneamente, aos vetores de entrada e aos alvos. Uma primeira avaliação considerou os desvios angulares medidos, em graus, entre as componentes extraídas e as de referência. Em razão do elevado número de componentes, para facilitar a visualização, esta análise considerou faixas de 50 componentes, sendo avaliados os valores médios e o maior desvio por faixa. As faixas foram organizadas na seguinte forma: a abscissa 1 responde pela faixa de 1 a 50 componentes; a abscissa 51, pela faixa de 51 a 100, e assim, suscessivamente. Este procedimento foi utilizado para a maior parte das análises deste capítulo. Como os valores de ângulo podem ser positivos e negativos, o valor médio foi calculado como a raiz do valor médio do quadrado dos valores de cada faixa. Na Figura 3.11(a) são exibidos os valores médios por faixa de componentes; enquanto na Figura 3.11(b), o maior valor verificado em cada faixa (em módulo). Os métodos estão indicados na seguinte forma: POT (potências), RPROP e CG (gradiente conjugado). Os valores médio e máximo dos desvios angulares para algumas faixas de componentes são apresentados na Tabela 3.6.

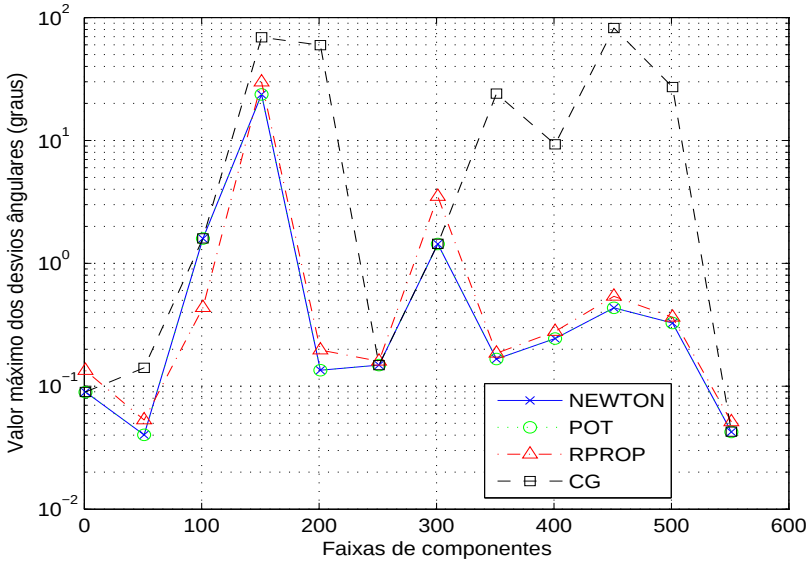
Tabela 3.6: Valores médios e máximos dos desvios angulares para os diferentes métodos de extração

	Valor médio				**Valor máximo**			
Faixa	**NEW**	**POT**	**RPROP**	**CG**	**NEW**	**POT**	**RPROP**	**CG**
$\mathbf{1-50}$	$0,019$	$0,019$	$0,028$	$0,022$	$0,09$	$0,089$	$0,13$	$0,09$
$\mathbf{100-150}$	$0,33$	$0,32$	$0,19$	$0,33$	$1,6$	$1,6$	$0,43$	$1,6$
$\mathbf{200-250}$	$0,11$	$0,11$	$0,12$	12	$0,13$	$0,13$	$0,19$	$59,6$
$\mathbf{300-350}$	$0,34$	$0,34$	$0,79$	$0,35$	$1,4$	$1,4$	$3,5$	$1,4$
$\mathbf{400-450}$	$0,12$	$0,12$	$0,14$	$1,9$	$0,24$	$0,24$	$0,27$	$9,3$

É possível perceber que a faixa de 151 a 200 é a mais crítica para quase



(a)



(b)

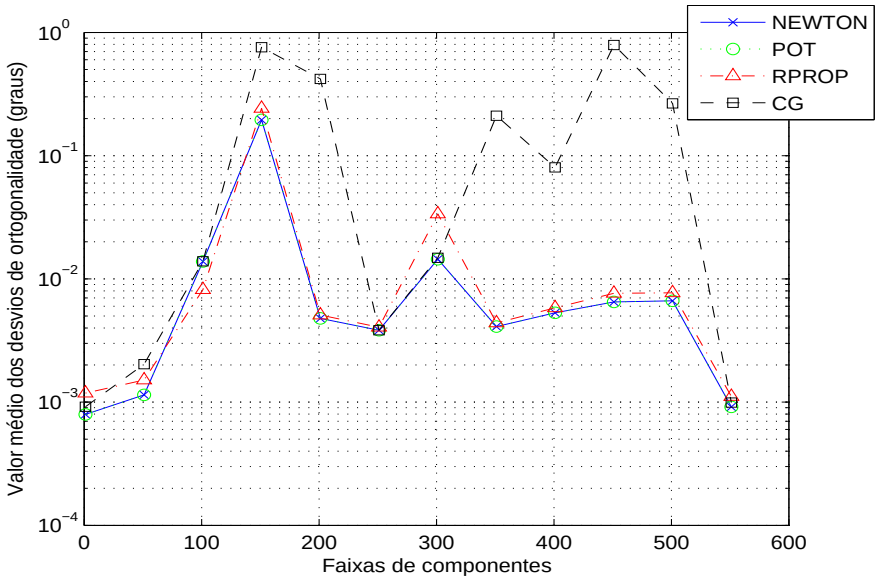
Figura 3.11: Valores médios (a) e máximos (b) dos desvios ângulares por faixa de componente para os diferentes métodos de extração.

todos métodos. O pior desempenho desta faixa é obtido para a componente 191 nos métodos NEWTON, POT e RPROP; e 198, para o método CG. Os valores dos desvios angulares, nos piores casos, são de $23,7^o$ (NEWTON,POT), $29,9^o$ (RPROP) e $69,1^o$ (CG). Os métodos de melhor desempenho foram o NEWTON e POT, com um comportamento bastante similar. Em segundo lugar, tem-se o RPROP e, por último, o CG. Excetuando-se o CG, os algoritmos apresentaram uma boa acurácia para a maior parte das direções extraídas, apresentando, exceto para a faixa mais crítica, desvios angulares inferiores a 2^o, no pior caso.

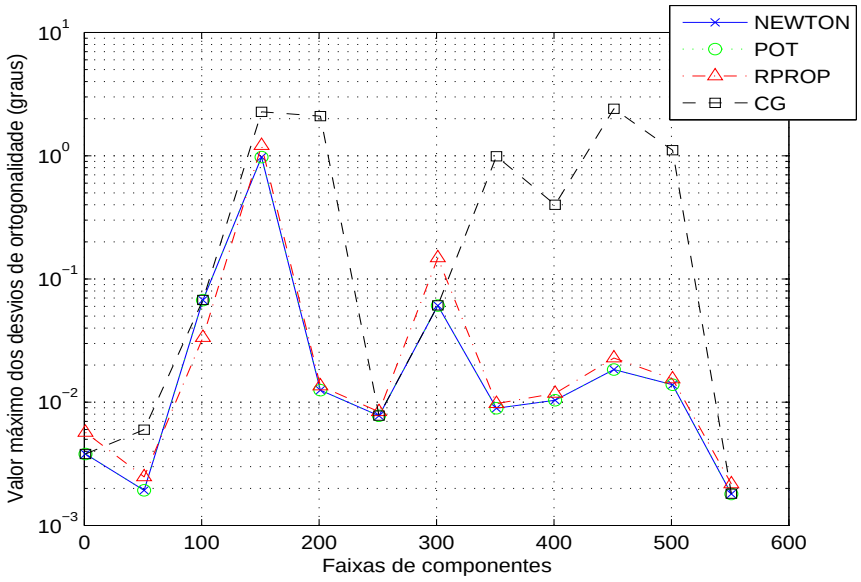
O segundo parâmetro avaliado foi a ortogonalidade entre as componentes extraídas. Como a ortogonalidade está diretamente relacionada à acurácia na estimação das direções, é esperado um comportamento análogo à análise anterior. Os resultados desta análise são apresentados na Figura 3.12. De forma coerente, tem-se um resultado similar entre as técnicas NEWTON e POT, as melhores, que são seguidas pelo algoritmo RPROP e, por último, pelo CG.

A terceira análise contemplou a acuidade dos autovalores, que também está relacionada à acuidade das componentes estimadas. Na Figura 3.13 são apresentados os resultados desta análise. Interessante observar que as técnicas NEWTON, POT e RPROP possuem, no pior caso, um erro relativo de 10^{-4}, obtido para a faixa de 151 a 200. Para a técnica CG, o erro é significativamente maior, inferior, para o pior caso, ao valor de 10^{-2}. Em termos dos valores médios, para o pior caso, o erro é, aproximadamente, uma ordem de grandeza menor que os valores máximos, da ordem de 10^{-5} para as técnicas NEWTON, POT e RPROP, e de 10^{-3}, para a CG.

Por fim, foi realizada uma comparação quanto ao número de passos requeridos para que o algoritmo atinja a convergência. Na Figura 3.14 e na Tabela 3.7 são apresentados os valores médios (arredondados para um valor inteiro) e máximos do número de passos até a convergência para algumas faixas de componentes. Pode-se perceber que todos os métodos apresentam uma mesma ordem de grandeza quanto ao número de passos. Os métodos NEWTON, POT e CG apresentam valores médios bastante similares, enquanto que, para o RPROP, os valores médios são pouco superiores aos primeiros, para a maior parte das faixas contempladas no gráfico. Considerando ainda os valores médios por faixa, tem-se, em média, de 1000 a 3000 passos até a convergência.

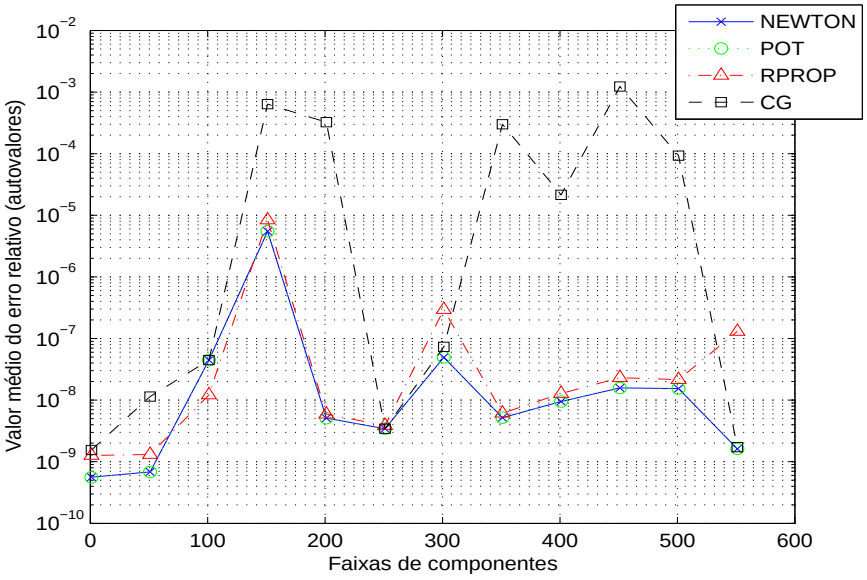


(a)

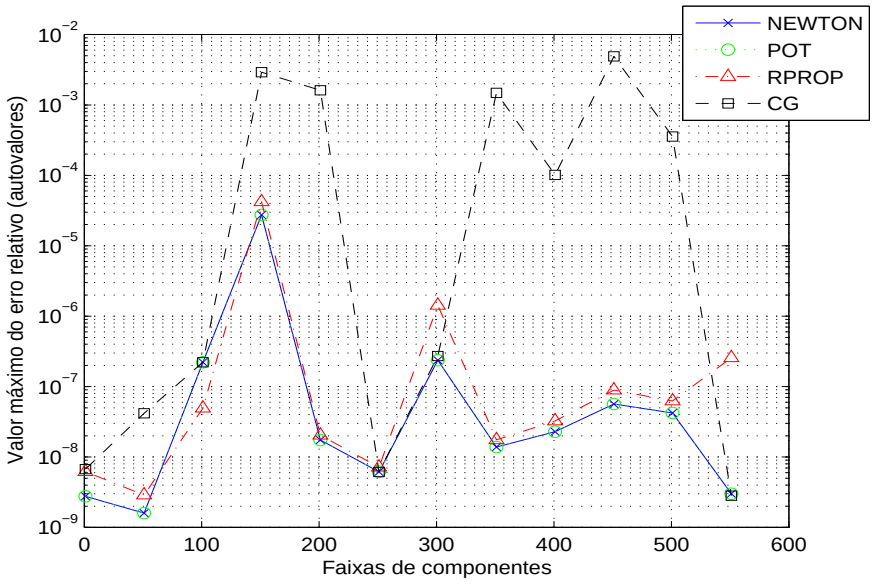


(b)

Figura 3.12: Valores médios (a) e máximos (b) dos desvios de ortogonalidade para as componentes extraídas pelos diferentes métodos de extração

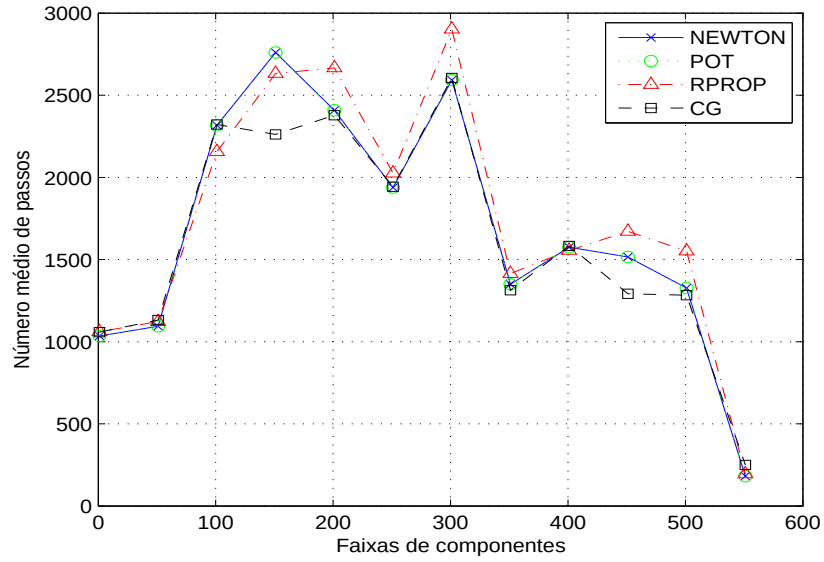


(a)

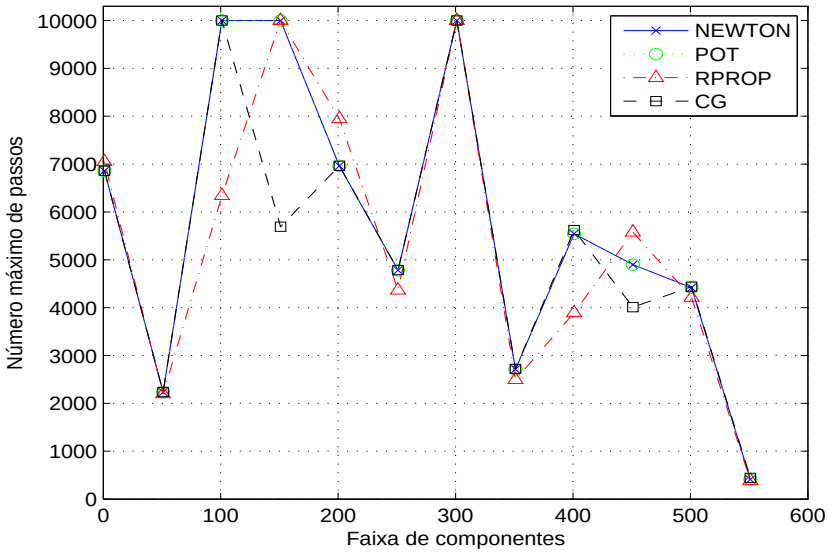


(b)

Figura 3.13: Acuidade na estimação dos autovalores (valor médio (a) e máximo (b)) por faixa de componente para os diferentes métodos de extração



(a)



(b)

Figura 3.14: Valores médios (a) e máximos (b) do número de passos para os diferentes métodos de extração

Conclui-se que dos métodos considerados nesta análise, os que conjugam um melhor compromisso entre a acuidade e o custo computacional são o POT e NEWTON, o último com custo computacional menor por iteração. Por seu turno, o RPROP tem um desempenho próximo a estes.

Tabela 3.7: Valores médios e máximos do número de passos até convergência para os diferentes métodos

	Valor médio				Valor máximo			
Faixa	**NEW**	**POT**	**RPROP**	**CG**	**NEW**	**POT**	**RPROP**	**CG**
$\mathbf{1-50}$	1035	1035	1061	1059	6857	6858	7049	6861
$\mathbf{100-150}$	2317	2317	2157	2322	10000	10000	6339	10000
$\mathbf{200-250}$	2409	2409	2667	2378	6961	6961	7949	6966
$\mathbf{300-350}$	2593	2593	2903	2606	10000	10000	10000	10000
$\mathbf{400-450}$	1576	1576	1556	1583	5544	5544	3894	5622

3.10.2 Comparação entre o método de Newton e os métodos GHA e PASTd

Nesta análise foram comparados o desempenho do método de NEWTON com os algoritmos GHA e PASTd, amplamente referidos na literatura. Conforme discussão anterior, para o método GHA deve-se definir o valor da constante η, e para o PASTd, o valor β. Vários ensaios foram realizados, para diferentes valores de η e β. Para η, mantido fixo por todo treinamento, foi verificado que sua redução produzia uma extração de maior acuidade, apresentando, porém, um número maior de passos até a convergência. Tendência similar foi verificada para o parâmetro β no algoritmo PASTd, quando seu valor se aproximava de um. Destes ensaios, os valores que apresentaram um melhor compromisso entre a acuidade e o número de passos até a convergência foram: $\eta = 1,5E-4$ e $\beta = 0,999995$, os quais foram adotados nas simulações consideradas a seguir. Em geral, na literatura, estas constantes, que são dependentes do problema, são arbitradas ou escolhidas por tentativa e erro.

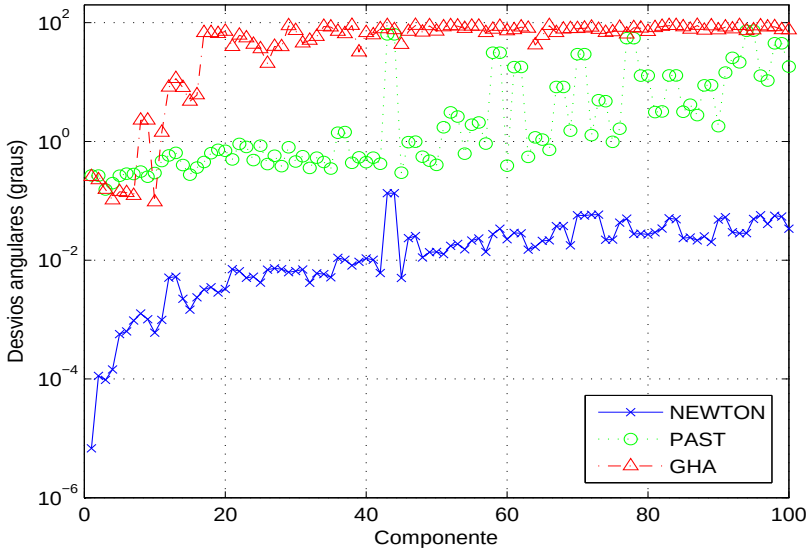
Na Figura 3.15 são apresentados os desvios angulares e de ortogonalidade da $1^{a.}$ a $100^{a.}$ componentes. Alguns valores para os desvios angulares são apresentados na Tabela 3.8. É possível perceber que desvios angulares elevados começam

a ocorrer a partir da $17^{a.}$ componente $(67,8^o)$, para o algoritmo GHA; e, para a $42^{a.}$ $(63,6^o)$, considerando o método PASTd. Pode-se perceber que o método de NEWTON é, no mínimo, uma ordem de grandeza mais acurado que o PASTd. O desempenho do PASTd é significativamente melhor que o GHA, neste aspecto, na faixa em consideração. Caso fosse arbitrado um limite de desvio angular de 45^o para as componentes extraídas, os métodos GHA e PAST teriam sido capazes de extrair apenas 16 e 41 componentes, respectivamente. Quanto a ortogonalidade, fato se repete, sendo o método de NEWTON de uma a duas ordens de grandeza mais acurado que os demais.

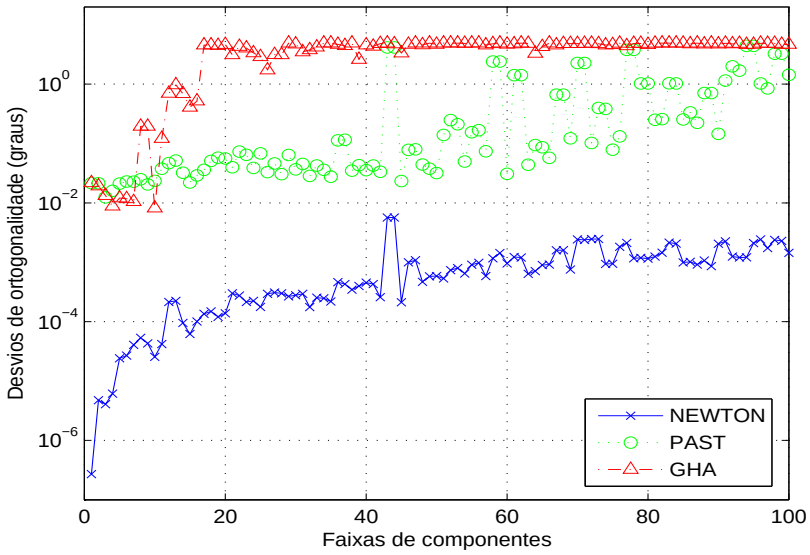
Tabela 3.8: Desvios angulares por componente extraída para os diferentes métodos

Componente	**GHA**	**PAST**	**NEW**
1	$0,26$	$0,27$	$3,4E-6$
15	$4,7$	$0,28$	$1,0E-3$
16	$6,0$	$0,36$	$1,7E-3$
17	$67,8$	$0,45$	$3,2E-3$
30	$73,4$	$0,46$	$5,3E-3$
42	$89,4$	$63,6$	$0,13$
43	$75,5$	$63,6$	$0,13$

A acuidade dos autovalores e o número de passos necessários para atingir a convergência são apresentados nas Figuras 3.16(a) e 3.16(b), respectivamente. Na Tabela 3.9 são apresentados, para algumas componentes, o número de passos para a convergência. Foi considerado que um passo para os algoritmos GHA e PASTd corresponde a uma passagem de todo conjunto de dados, ou seja, a N atualizações de pesos, onde N é o número de eventos disponíveis. Assim, para os três métodos em análise, é possível realizar uma comparação do seu custo computacional para a extração através da avaliação do número de passos até a convergência, uma vez que, segundo esta contagem, os métodos passam a apresentar um custo computacional por passo equivalente. Verifica-se que o método de NEWTON é quase 4 ordens de grandeza mais acurado na estimação dos autovalores que os demais na faixa considerada. Com respeito ao número de passos, os algoritmos GHA e PAST apresentaram valores, em média, de 20 a 40 vezes maiores que o do método de NEWTON. Pode-se



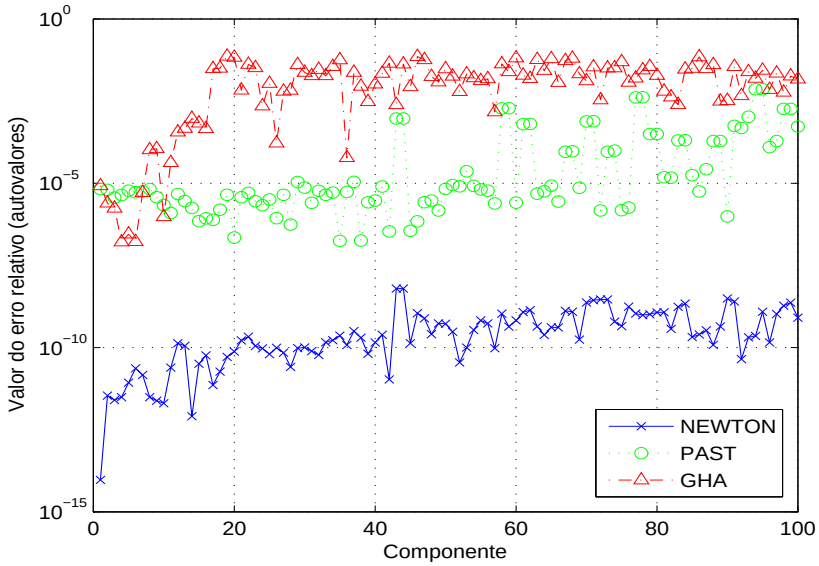
(a)



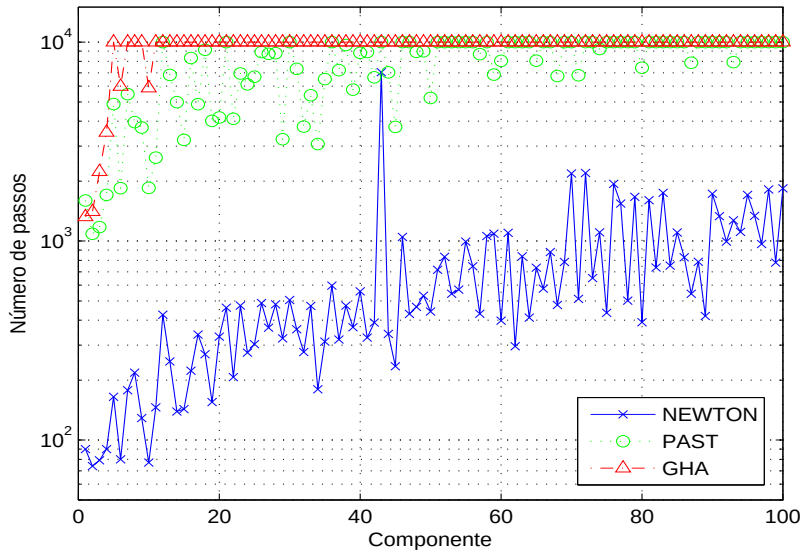
(b)

Figura 3.15: Valores dos desvios angulares (a) e dos desvios de ortogonalidade por componente extraída (b) para os diferentes métodos.

verificar ainda que para grande número de componentes, os algoritmos GHA e PAST atingiram o valor limite de 10000 passos, valor para o qual a extração é interrompida. Dos resultados, o método de melhor desempenho é o NEWTON, seguido pelo PASTd e, por último, pelo GHA.



(a)



(b)

Figura 3.16: Acuidade na estimação dos autovalores (a) e número de passos (b) por componente extraída para os diferentes métodos

Tabela 3.9: Passos por componente extraída para os diferentes métodos de extração

	Número de passos		
Componente	**GHA**	**PASTd**	**NEWTON**
1	1223	1599	47
15	10000	3221	129
16	10000	8314	211
17	10000	4871	339
30	10000	10000	478
42	10000	10000	1119
43	10000	10000	430

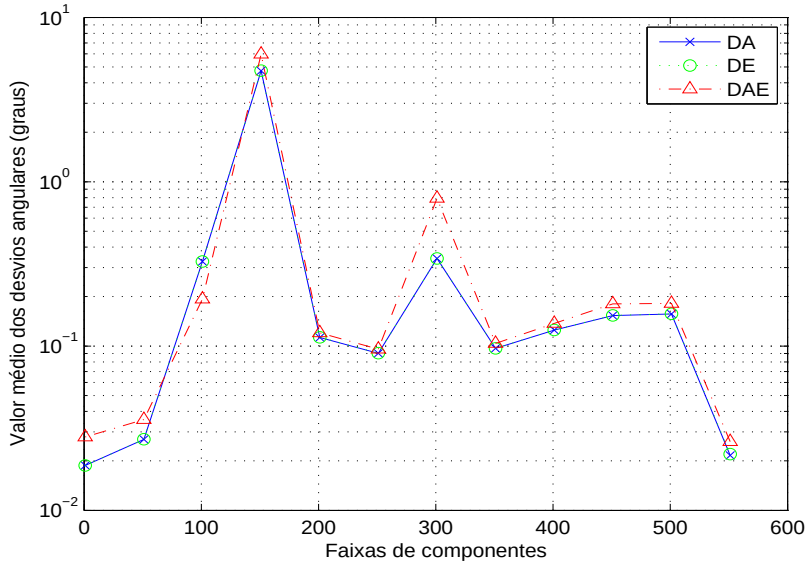
3.10.3 Comparação entre as diferentes modalidades de deflação

A última análise visou comparar o desempenho das diferentes modalidade de deflação, considerando o método de NEWTON, para a extração das componentes. Os valores dos desvios angulares médios e máximos para posicionamento tabela cada faixa de componentes são exibidos nas Figuras 3.17(a) e 3.17(b), respectivamente. Alguns destes valores são apresentados na Tabela 3.10. Pode-se perceber que as modalidades DA (deflação aos alvos) e DE (deflação às entradas) possuem desempenho praticamente equivalente em relação à acuidade das componentes extraídas, enquanto o desempenho da modalidade DAE (deflação às entradas e alvos) é ligeiramente inferior ao apresentado pelas duas primeiras modalidades.

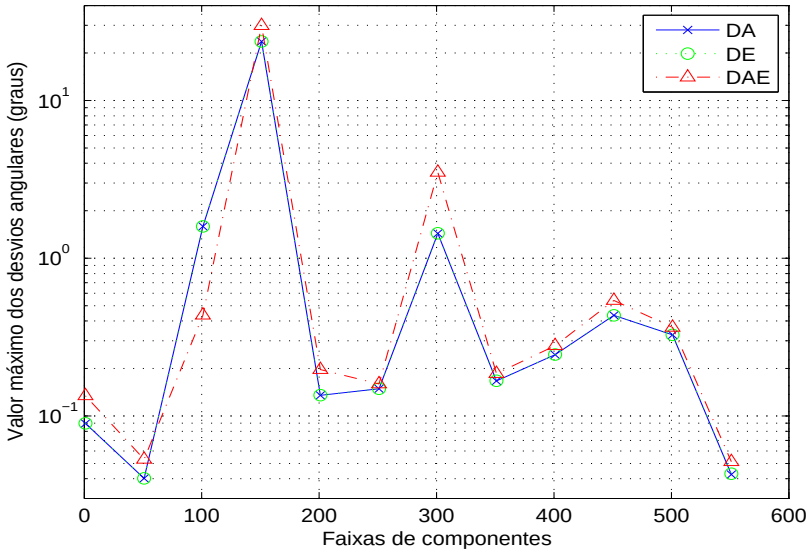
Quanto à ortogonalidade, conforme as Figuras 3.18(a) e 3.18(b), e a acuidade dos autovalores estimados com base nas componentes extraídas, conforme as Figuras 3.19(a) e 3.19(b), verifica-se um desempenho equivalente das três modalidades de deflação, que é um reflexo do comportamento verificado para a direção das componentes extraídas.

Com relação ao número de passos, conforme as Figuras 3.20(a) e 3.20(b) e a Tabela 3.11, as modalidades DA e DE apresentam comportamento similar, enquanto a DAE, em média, apresenta valores iguais ou superiores às demais modalidades.

Conclui-se que, com respeito à deflação, as modalidades DA e DE possuem desempenho similar. Um desempenho ligeiramente inferior é obtido pela DAE, ainda

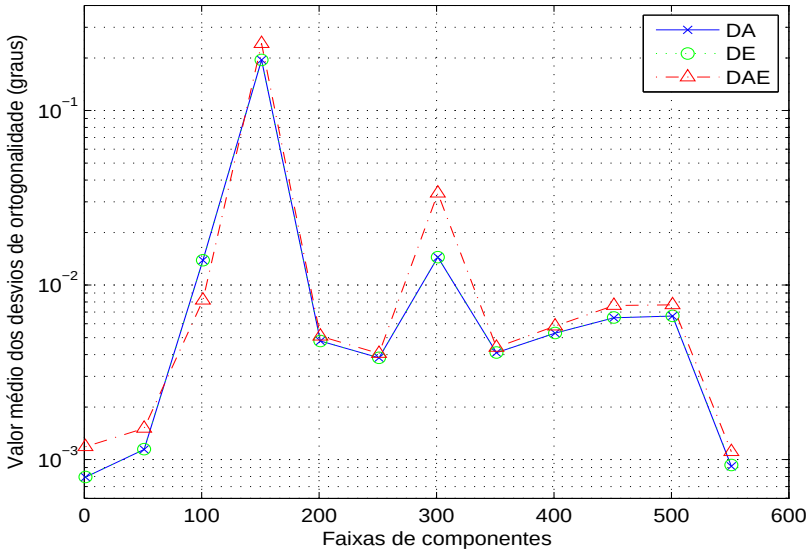


(a)

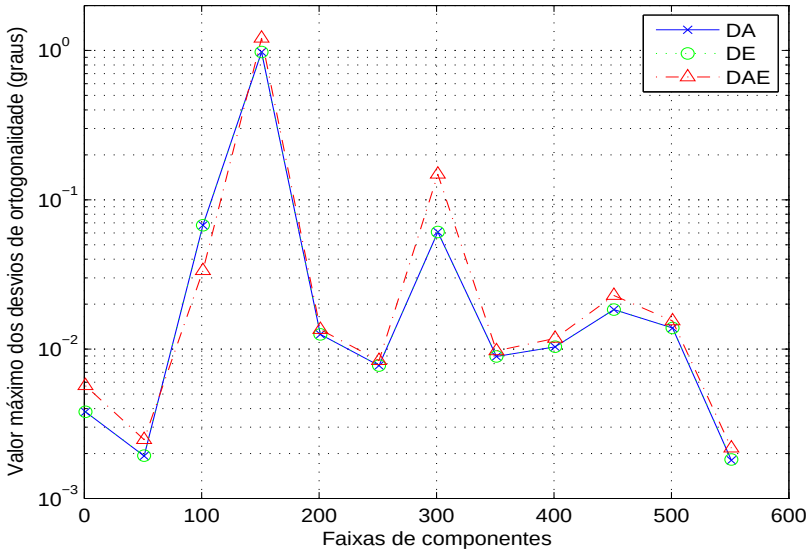


(b)

Figura 3.17: Valores médios (a) e máximos (b) dos desvios angulares por faixa de componentes para os diferentes métodos de extração

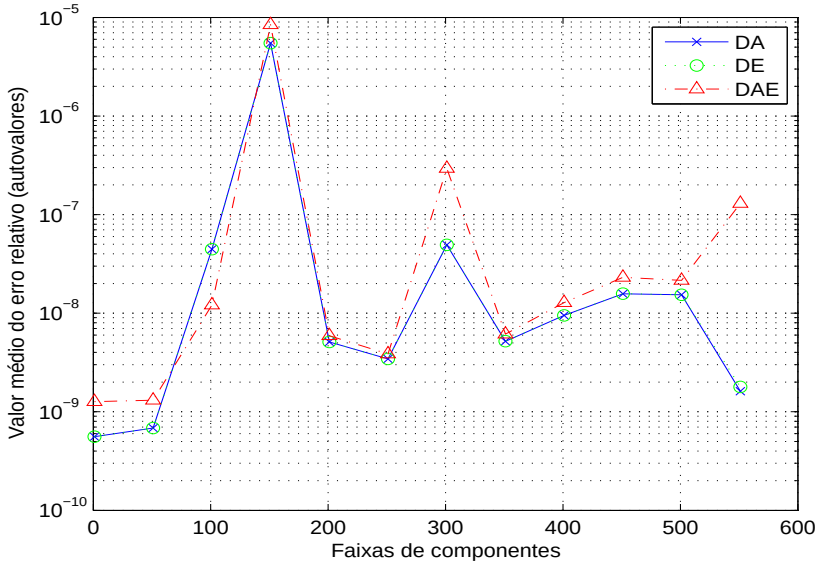


(a)

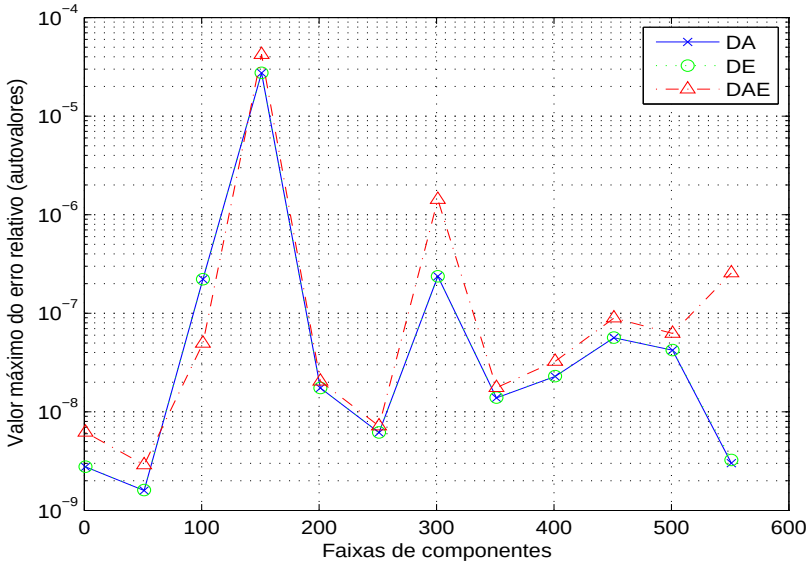


(b)

Figura 3.18: Valores médios (a) e máximos (b) dos desvios de ortogonalidade por faixa de componente para as diferentes modalidades de deflação



(a)



(b)

Figura 3.19: Acuidade na estimação dos autovalores (valor médio (a) e máximo (b)) por faixa de componente para as diferentes modalidades de deflação

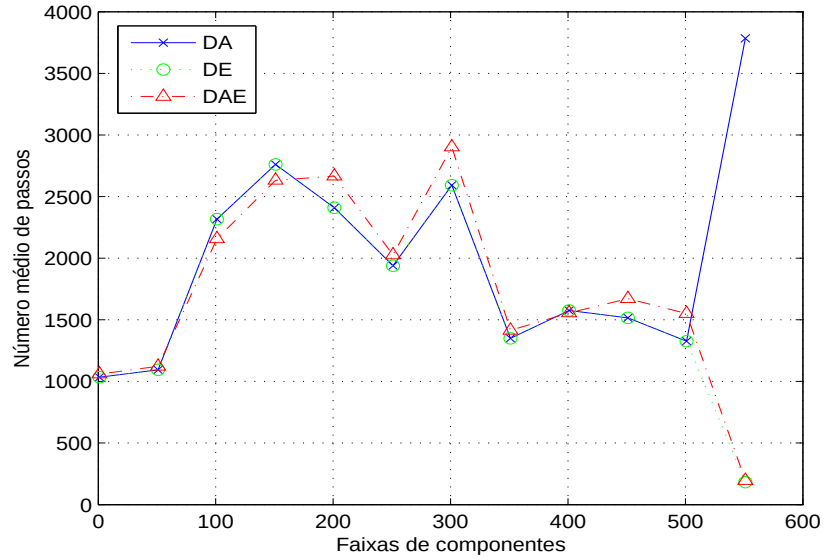
Tabela 3.10: Valores médios e máximos para os desvios angulares por faixa de componentes para as diferentes modalidades de deflação

	Valor médio			**Valor máximo**		
Faixa	**DA**	**DE**	**DAE**	**DA**	**DE**	**DAE**
1 – 50	0,019	0,019	0,028	0,089	0,089	0,13
100 – 150	0,32	0,32	0,19	1,6	1,6	0,43
200 – 250	0,11	0,11	0,12	0,13	0,14	0,20
300 – 350	0,34	0,34	0,79	1,4	1,4	3,5
400 – 450	0,13	0,13	0,14	0,24	0,25	0,28

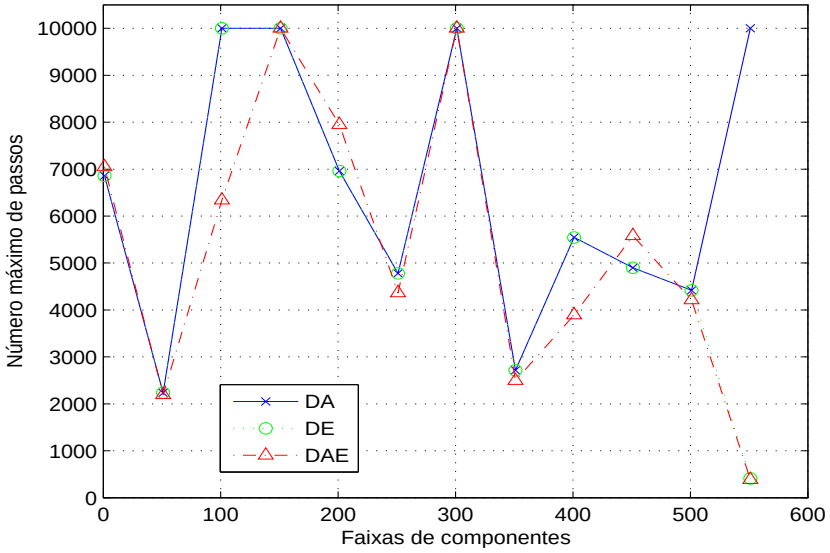
que a diferença não seja muito significativa. Esta diferença se deve ao ruído introduzido no sistema de extração pela transformação de deflação, que é maior na modalidade DAE, já que afeta tanto os alvos quanto as entradas.

Tabela 3.11: Valores médios e máximos do número de passos para as diferentes modalidades de deflação

	Número médio			**Número máximo**		
Faixa	**DA**	**DE**	**DAE**	**DA**	**DE**	**DAE**
1 – 50	1035	1035	1061	6858	6858	7049
100 – 150	2317	2317	2157	10000	10000	6339
200 – 250	2409	2409	2667	6961	6961	7949
300 – 350	2593	2593	2903	10000	10000	10000
400 – 450	1576	1576	1556	5545	5544	3894



(a)



(b)

Figura 3.20: Valores médios (a) e máximos (b) dos número de passos até a convergência por faixa de componentes para as diferentes modalidades de deflação

Capítulo 4

Projeto do classificador neural

Num sistema de classificação automática de contatos, há requisitos severos quanto à eficácia e à confiabilidade. Conforme discutido no capítulo 2, em razão da complexidade do problema e da pluraridade de cenários possíveis, a capacidade de generalização é um parâmetro crítico, e restrições estatísticas quanto à caracterização das classes são esperadas.

Neste contexto, o desenvolvimento de classificadores eficientes demanda cuidados especiais quanto ao projeto do classificador, o que se traduz numa escolha apropriada da topologia, do algoritmo de treinamento e dos dados utilizados para seu desenvolvimento e avaliação. Numa aplicação crítica como o sonar, uma previsão apropriada do desempenho do classificador em cenários de operação faz-se necessária, a qual exige a definição de índices e técnicas para a estimação do seu desempenho. É também útil que uma caracterização estatística dos dados disponíveis seja realizada, identificando-se quais classes possuem uma classificação mais crítica e/ou maiores restrições quanto à caracterização estatística. Para as classes identificadas como "problemáticas", um maior cuidado pode ser dedicado pelo operador na tomada de decisões, e a necessidade de aquisições complementares pode ser sinalizada.

Neste capítulo, são discutidos os principais parâmetros envolvidos no projeto de classificadores neurais, em especial aqueles que possuem impacto na sua capaciadade de generalização. Para viabilizar sua avaliação, são discutidos também índices e técnicas para a estimação de seu desempenho. Através de classificadores neurais MLP [4] totalmente conectados, é realizada uma caracterização estatística do

conjunto de dados em estudo e previsto seu desempenho em diferentes cenários de operação.

4.1 Classificadores neurais para sonar passivo

Característica relevante a qualquer sistema de classificação é a capacidade de generalização, isto é, de classificar corretamente eventos pertencentes a classes conhecidas que não tenham sido utilizados no seu treinamento. O desempenho de um sistema de classificação quanto à generalização está diretamente relacionado à qualidade de seu aprendizado, que é dependente de diferentes fatores, entre eles: o conteúdo estatístico dos dados, a forma como as informações relevantes são disponibilizadas nos eventos, a técnica de classificação e o algoritmo de treinamento selecionado, assim como o processo de controle do aprendizado.

Com respeito aos dados, como a extração das características relevantes à classificação, assim como as regras de decisão, serão neles baseadas, é desejável que o conjunto possua uma estatística suficiente ao problema, isto é, que contemple o maior número possível de cenários relevantes, em especial para as classes de classificação mais difícil. É ainda esperado, quantitativa e qualitativamente, que as especificidades das classes sejam apropriadamente representadas, evitando-se a especialização em casos particulares.

Aspecto importante no desempenho de um classificador neural é como a informação relevante à classificação está codificada nos dados, ou seja, quais variáveis estão diretamente relacionadas com a caracterização do problema de interesse. Freqüentemente, em diferentes aplicações, os dados são provenientes de espaços de dimensão elevada e, eventualmente, a maior parte das componentes é irrelavante à solução do problema. Reduzir o número de variáveis à entrada de um classificador, eliminando as variáveis descorrelacionadas ao problema, é um procedimento útil, pois simplifica o processo de aprendizado, e conduz, freqüentemente, a uma redução do custo computacional envolvido, o que é atraente para classificadores que operam em tempo real, tais como o de sonar passivo. A redução do número de variáveis permite ainda ao classificador melhor discernir sobre os dados, o que resulta em melhor desempenho. Assim, de forma usual, sistemas de classificação exploram o

pré-processamento dos dados para enfatizar e/ou selecionar as informações relevantes ao problema.

A escolha da técnica de classificação está relacionada às potencialidades do classificador, as quais, para a produção de um classificador eficiente, devem ser apropriadamente exploradas pelo algoritmo de aprendizado. Adicionalmente, cabe ao controle de aprendizado regular o treinamento, evitando uma demasiada especialização nos dados, a qual pode comprometer sua capacidade de generalização.

Produzir um classificador com base num dado conjunto de dados exige, usualmente, sua partição em duas amostras, referidas como conjuntos de projeto e de avaliação (teste): o primeiro é utilizado para o treinamento do classsificador; e o segundo, para inferir sua capacidade de generalização. Produzir e avaliar corretamente um classificador presume que ambas amostras possuam um conteúdo estatístico representativo do problema, sob o prejuízo de comprometer o aprendizado ou prover uma estimativa não-realística do seu desempenho. Assim, em condições de restrição quanto à caracterização estatística da base de dados, a escolha destes conjuntos deve ser criteriosa, sendo estabelecido um compromisso entre as qualidades de aprendizado e de avaliação.

Um parâmetro importante no projeto de um classicador neural MLP totalmente conectado é o número de camadas e a quantidade de neurônios por camada. Estes parâmetros estão diretamente relacionados à capacidade da rede em absorver as regras de classificação necessárias à solução do problema, refletindo-se no erro de aproximação do classificador [116]. Em geral, tanto mais complexo o problema, mais complexa é a topologia adequada para sua solução. Usualmente, classificadores MLP consideram topologias com 2 camadas [4], onde o número de nós de entrada é definido pela dimensão do espaço dos dados; e o número de neurônios de saída, pelo número de classes existentes, já que as classes são codificadas de forma maximamente esparsa [4]. Cabe ao projetista, portanto, uma seleção apropriada do número de neurônios da camada intermediária, a qual, normalmente, é dependente do problema em questão.

Aumentar a complexidade da rede acarreta, em geral, numa redução do erro de aproximação da rede, porém pode comprometer o processo de aprendizado, o qual passa a envolver a estimação de um maior número de parâmetros com base nas ca-

racterísticas existentes nos dados, afetando a capacidade de generalização. Modelos mais complexos requerem um maior número de graus de liberdadade, o que torna o processo de ajuste dos parâmetros necessários mais complexo e demanda, normalmente, um maior conteúdo estatístico dos dados. A quantidade de parâmetros existentes impacta no erro de estimação do classificador [116], o qual reflete a capacidade do algoritmo de aprendizado em extrair as regras necessárias à classificação. Assim, a escolha da topologia deve ser parcimoniosa [117], ou seja, o classificador não deve crescer além do necessário, sobre o risco de haver um *overfitting* [49] dos parâmetros, o que compromete a capacidade de generalização do modelo.

Outra fonte de *overfitting* é o número de épocas [4] do treinamento. Caso nenhuma restrição seja imposta ao longo do treinamento da rede, há uma tendência do classificador em "memorizar"os dados de treinamento, perdendo sua capacidade de generalização [4, 49, 118]. Assim, para problemas com restrições estatísticas, tais como o sonar passivo, o problema do *overfitting* é mais crítico, exigindo maiores cuidados no projeto do classificador. Na literatura são propostas várias alternativas para lidar com o problema do *overfitting*. Didaticamente, estas técnicas podem ser agrupadas em duas linhas: uma primeira, que utiliza todo conjunto de dados para a seleção da topologia, aprendizado e avaliação do classificador; e uma segunda, onde o conjunto de dados é particionado em subconjuntos, conforme detalhado a seguir.

Segundo a primeira linha, o controle do número de parâmetros envolvidos pode ser realizado através de técnicas de poda [119, 120, 121] ou regularização [122], ou seja, parte-se de um modelo maior, cuja complexidade é reduzida pela respectiva técnica. Em caminho inverso, algumas técnicas propõem a construção e a avaliação de modelos de complexidade crescente, sendo um dos modelos escolhido através de critérios heurísticos [123] ou através de testes de hipóteses [124, 125]. Para outros trabalhos, a escolha da topologia baseia-se em índices estatísticos [126, 127, 128]. Em todos os casos, não há um controle explícito do número de épocas do treinamento, o qual é realizado implicitamente, de acordo com a estratégia adotada por cada técnica.

A segunda linha contempla as técnicas baseadas em validação cruzada [129], que propõem a divisão dos dados em dois conjuntos: o de projeto e o de avaliação; o primeiro utilizado para a obtenção do modelo, e o segundo para estimar a sua

capacidade de generalização. A validação cruzada pode ser utilizada tanto para a escolha da complexidade do modelo quanto para a determinação da parada do treinamento. Para a definição do modelo, classificadores de diferentes topologias são construídos e avaliados, sendo escolhida a topologia de maior capacidade de generalização. Em relação à parada do treinamento, este é interrompido quando é sinalizada uma perda na capacidade de generalização, o que é referido pela literatura como parada antecipada [4].

No projeto de classificadores neurais baseados na técnica de validação cruzada, os dados são comumente partilhados em três conjuntos disjuntos: treino, validação e teste [49]. O conjunto de treino corresponde ao conjunto de projeto. Os conjuntos de validação e teste realizam o papel do conjunto de avaliação em duas etapas distintas: na parada antecipada (validação) e na avaliação final do classificador (teste). A validação cruzada não pressupõe nenhuma suposição estatística sobre os dados ou classificador [124], possui fácil implementação [130] e costuma apresentar desempenho superior que a regularização [131], o que motivou sua adoção no problema do sonar.

Na formação dos conjuntos de treino, teste e validação é relevante definir quantos eventos irão constituir cada conjunto. Tendo em vista a complexidade do problema e as prováveis restrições quanto à caracterização estatística das classes, ao considerar a partição dos dados disponíveis entre os três conjuntos, é atraente disponibilizar o maior número possível de eventos para os conjuntos de treino e teste, que estão diretamente relacionados com a qualidade do aprendizado e da avaliação. Aumentar o número de eventos utilizado nestes conjuntos resulta numa redução do número de eventos disponível ao conjunto de validação, o que pode prejudicar a parada do treinamento. Estabelecer um compromisso entre estes diferentes requisitos é uma tarefa dependente do problema, que não é simples. Na literatura não foram encontradas referências sobre este tema. Freqüentemente, trabalhos que exploram a técnica da validação cruzada utilizam conjuntos de tamanho arbitrário, sem prover uma discussão ou justificativa para as escolhas realizadas. Deste modo, optou-se por dividir o conjunto de dados em duas partições: a primeira, utilizada apenas para a obtenção do classificador; e a segunda, explorada tanto para a parada quanto para a avaliação do classificador, resultando em 2 conjuntos, aqui referidos como de projeto

e avaliação.

Outro aspecto relevante é a escolha do algoritmo de treinamento. A classificação de sinais de sonar passivo é um problema sabidamente complexo, o que exige do algoritmo de treinamento a habilidade de extrair a informação necessária a este fim, mesmo na existência de prováveis restrições quanto à caracterização estatística das classes. Outro agravante é o número e a dimensionalidade elevada dos sinais a serem analisados. Estes aspectos orientaram a escolha da técnica RPROP [110] para o treinamento de todos classificadores utilizados nesta tese, visto que esta técnica conjuga eficácia e baixo custo computacional por iteração, conforme discutido na seção 3.8.2.3. Os valores dos parâmetros utilizados para a técnica foram os sugeridos pela referência [110], considerando como função objetivo o erro quadrático médio para a derivação dos gradientes necessários.

4.2 Avaliação de classificadores neurais

Prover mecanismos de avaliação de desempenho para o sistema de classificação automática de contatos é de especial importância. Esta avaliação pode ser utilizada tanto para a seleção dos parâmetros envolvidos no seu projeto (conjunto de dados, topologia, parada, entre outros), quanto para caracterizar a base de dados disponível. No último caso, é possível identificar se o conteúdo estatístico disponível nos dados atende às exigências do problema, se há restrições estatísticas na caracterização das classes, quais delas são de classificação mais difícil e as incertezas de classificação envolvidas. Este conjunto de informações pode ser considerado pelo operador na tomada de decisão, assim como pode orientar novas aquisições de dados.

Para avaliar um classificador, faz-se necessário definir índices e técnicas de estimação de seu desempenho, as quais serão discutidas a seguir.

4.2.1 Índices de desempenho

Classificadores neurais usualmente utilizam uma codificação maximamente esparsa das classes [4], onde cada neurônio da saída da rede é atribuído a uma classe e treinado para apresentar saída +1, em caso de eventos pertencentes a sua classe; e -1, para os eventos pertencentes as demais classes. Decidir, no entanto, em favor

de uma classe exige um critério, referido como de decisão, que defina com base nas saídas da rede o contato mais provável.

Em problemas binários de classificação, a decisão é baseada na comparação da saída da rede com um limiar que pode ser definido através de diferentes critérios [27, 52]. Para classificadores de M classes, o problema é mais complexo e demanda a definição de um conjunto de regras de decisão [27]. Em ambos os casos, o limiar ou as regras necessárias à classificação podem ser obtidos pela otimização de uma função de custo [27] ou perda [1]. Seja $P(C_i|C_j)$ a probabilidade de um evento da classe C_j ser classificado como pertencente à classe C_i. O custo médio associado às decisões tomadas por um classificador arbitrário pode ser expresso como [27]:

$$c = \sum_{i=1}^{M} \lambda_{ii} P(C_i) P(C_i|C_i) + \sum_{i=1}^{M} \sum_{j \neq i, i=1}^{M} \lambda_{ij} P(C_i) P(C_j|C_i), \tag{4.1}$$

onde λ_{ij} responde pelo custo da identificação de eventos da classe i como pertencentes à classe j. No primeiro termo, contabilizam-se os acertos na classificação das M classes; no segundo, os diferentes erros de classificação possíveis.

Definir através da Equação 4.1 um conjunto ótimo de regras de decisão exige um conhecimento a priori da probabilidade de ocorrência de cada classe, assim como dos riscos relacionados a cada uma das M^2 decisões possíveis. Por restrições operacionais e questões de segurança nacional, o IPqM não disponibilizou nenhum destes parâmetros para o conjunto de dados utilizado neste trabalho. Na inexistência destes parâmetros, suposição comum é considerar as classes equiprováveis e os custos das decisões corretas e incorretas como 0 e 1, respectivamente, ainda que, num cenário real, a importância e a probabilidade de ocorrência das diferentes classes de navios seja distinta. Através desta suposições, a decisão reduz-se ao simples critério de máxima probabilidade a posteriori (MAP) [27], que, aplicado a classificadores neurais, atribui ao evento a classe correspondente ao neurônio de maior valor de saída [4].

Definido o critério de decisão, a qualidade de um classificador pode ser avaliada através da estimação de sua eficiência de classificação, obtida através da análise de sua resposta para um conjunto de N vetores de teste independentes. Sejam e_i e N_i a probabilidade de acerto e o número de eventos da i-ésima classe. A probabilidade de k_i vetores serem corretamente classificados é dada pela distribuição

binomial [1]:

$$p = \binom{N_i}{k_i} e_i^{k_i} (1 - e_i)^{N_i - k_i}, \tag{4.2}$$

onde e_i é desconhecido.

Derivando a Equação 4.2 e igualando a zero, resulta que um estimador de máxima verossimilhança [132] para e_i é dado por:

$$\hat{e}_i = \frac{k_i}{N_i}, \tag{4.3}$$

logo pode ser determinado pela razão entre o número de eventos corretamente identificados pelo número total de eventos disponíveis para a classe. Cabe notar que $\hat{e}_i$ é um estimador não-tendencioso [52] da probabilidade real (e_i) de acerto ou da eficiência de detecção do classificador, pois [1]:

$$E[\hat{e}_i] = \frac{E[k_i]}{N_i} = e_i \tag{4.4}$$

O estimador $\hat{e}_i$ provê uma medida da eficiência real (e_i) do classificador, possuindo uma incerteza estatística associada. Uma forma possível de quantificar esta incerteza é através de sua variância, comumente indicada através do valor RMS (*Root Mean Square*) da medida, que pode ser determinado através de:

$$RMS^2_{\hat{e}_i} = \frac{\sigma^2_{k_i}}{N_i^2} = \frac{e_i(1 - e_i)}{N_i} \approx \frac{\hat{e}_i(1 - \hat{e}_i)}{N_i}, \tag{4.5}$$

que utilizou a estimativa provida por $\hat{e}_i$ como o valor da eficiência real de detecção (e_i).

Desconhecida a importância relativa de cada classe, uma medida intuitiva da qualidade do classificador é a eficiência média, que pode ser determinada por:

$$\hat{e}_M = \frac{1}{M} \sum_{i=1}^{M} \hat{e}_i, \tag{4.6}$$

e cuja variância vale:

$$\sigma^2_{\hat{e}_M} = \frac{1}{(M-1)^2} \sum_{i=1}^{M} \sigma^2_{\hat{e}_i}, \tag{4.7}$$

onde M corresponde ao número de classes.

[1] Nas deduções a seguir, duas propriedades da distribuição binomial serão utilizadas: $E[k_i] = N_i e_i$ e $\sigma^2_{k_i} = N_i e_i (1 - e_i)$ [12].

A seleção por eficiência média é simples, porém possui alguns inconvenientes. Um deles é que o desempenho ruim em uma classe pode ser mascarado por um bom desempenho nas demais. Numa aplicação de sonar, ainda que uma eficiência média elevada seja desejável, é relevante que todas as classes possuam valores de eficiência superiores a um dado patamar. Adicionalmente, como não há conhecimento sobre a importância relativa de cada classe, é desejável que o classificador possua valores de eficiência equilibrados para as classes. É interessante, portanto, considerar um índice que leve em consideração as similaridades entre as eficiências de detecção das diferentes classes.

Um critério que atende estes objetivos é o soma-produto (SP) [133]. Este critério se baseia na composição da média aritmética (e_M) com a média geométrica (e_G) das eficiências de cada classe segundo a relação:

$$e_{SP} = \sqrt{e_M e_G}, \tag{4.8}$$

e pode ser estimado utilizando:

$$\hat{e}_M = \frac{1}{M}\sum_{i=1}^{M}\hat{e}_i \tag{4.9}$$

e:

$$\hat{e}_G = \left(\prod_{i=1}^{M}\hat{e}_i\right)^{\frac{1}{M}} \tag{4.10}$$

A variância do estimador de eficiência soma-produto pode ser deduzido considerando que as eficiências de cada classe são variáveis aleatórias independentes, o que torna válida a seguinte relação [134]:

$$\sigma^2_{\hat{e}_{SP}} = \sum_{i=1}^{M}\left(\frac{\partial e_{SP}}{\partial e_i}\right)^2\sigma^2_{\hat{e}_i} \tag{4.11}$$

É fácil mostrar que o valor de $\frac{\partial e_{SP}}{\partial e_i}$ é dado por:

$$\frac{\partial e_{SP}}{\partial e_i} = \frac{1}{2M\hat{e}_M\hat{e}_G}\left(\frac{1}{\hat{e}_i} + \frac{1}{\hat{e}_M}\right), \tag{4.12}$$

onde M é o número de classes.

A eficiência média e SP são indicadores de desempenho de fácil interpretação e produção, porém possuem alguns inconvenientes. Um deles é a dependência do conjunto de regras de decisão que é aplicada à saída do classificador para a definição

de a qual classe o evento pertence, cuja escolha inapropriada pode comprometer o processo de avaliação. Dois classificadores de mesma eficiência podem ainda realizar uma classificação com níveis distintos de qualidade e confiabilidade. Exemplificando: suponha um problema de 2 classes, onde dois classificadores MLP são treinados para apresentar o valor $+1,0$ para os eventos da classe; e $-1,0$, para os eventos da não-classe. Considere ainda que o limiar de decisão é zero. Na saída de cada classificador são verificadas duas distribuições estatísticas, uma relacionada à classe, a qual se espera concentrar em torno de $+1$, e outra relacionada a não-classe, tipicamente, concentrada em -1. Ainda que ambos classificadores possam apresentar eficiências idênticas para a identificação de eventos da classe e não-classe, o formato das distribuições das saídas correspondentes pode ser completamente distinto. Quanto mais separadas estas distribuições, melhor será a distinção entre classe e não-classe, assim como a qualidade da classificação.

Outra forma de avaliar a qualidade de classificadores binários é através da curva de operação do receptor (*Receiver operating characteristic* - ROC) [27, 52]. A curva ROC esboça a probabilidade de detecção, isto é, de que eventos da classe sejam corretamente identificados, como função da probabilidade de falso-alarme, ou seja, de que os eventos da não-classe sejam considerados como pertencentes à classe. Para esboçar esta curva, varia-se o limiar de decisão dentro de uma faixa, identificando-se, para cada limiar, o valor de ambas probabilidades. A curva ROC reflete a separação das distribuições das saídas de um classificador para eventos da classe e não-classe, ou seja, classificadores cujas distribuições das saídas sejam mais separadas para classe e não-classe tendem a possuir curvas de crescimento mais rápido e acentuado. Quanto mais acentuado for o crescimento da curva, maior é o valor de detecção associado a um mesmo valor de falso-alarme, logo melhor é o classificador. Um índice comumente utilizado para qualificar uma curva ROC é a sua área, tanto maior quanto mais acentuado for seu crescimento [135], provendo uma medida que é independente da escolha do limiar de decisão.

A obtenção de um índice aplicável a classificadores de múltiplas classes que possua propriedades similares a área ROC é um tópico ainda em pesquisa. As propriedades geométricas da ROC de M-classes podem ser encontradas em [136]. Em [137] é mostrado que o cálculo do volume sobre a hipersuperfície associada a ROC

de M-classes não é um índice útil, a despeito de ser uma extensão intuitiva do conceito de área utilizado para 2 classes. Há indicativos ainda que a construção desta hipersuperficie para a determinação do valor ótimo da função de custo é computacionalmente inviável, tendo sido propostos critérios alternativos baseados em soluções sub-ótimas, tais como em [138].

Uma extensão simples da área da ROC para múltiplas classes consiste em reduzir o problema da detecção de M classes à detecção de duas classes. Nesta redução, dois enfoques são possíveis para a constituição dos conjuntos de classe e não-classe: uma classe contra todas [139] ou uma contra a outra [140]. Pelo primeiro critério, são construídas M ROCs, uma por classe, que consideram os eventos de todas as demais classes na formação do conjunto relativo à não-classe. No segundo são realizadas $M(M-1)$ ROCs que compreendem todas as comparações 2 a 2 possíveis entre classe e não-classe. Em ambos casos, para cada comparação, é calculada a área da ROC correspondente, e o índice final é calculado com base na média das áreas obtidas.

A avaliação de classificadores pela área da curva ROC esbarra no problema prático da estimação da curva e do valor de sua área. Freqüentemente, a área da ROC é estimada pela integração numérica da curva experimental, procedimento que pode não ser acurado o suficiente para dadas aplicações. Em aplicações mais críticas, pode ser necessário [135] propor e ajustar um modelo matemático à curva ROC, o que pode não ser trivial, ou mesmo, possível.

4.2.2 Estimação de desempenho

O projeto de um classificador demanda a existência ou obtenção de um conjunto de dados representativo do problema. Como, do ponto de vista prático, é impossível dispor de toda população dos dados, a qual teria um número infinito de exemplares, qualquer conjunto de dados é uma amostra, ou seja, representa um subconjunto finito desta população. O desempenho do classificador está, portanto, sujeito ao processo de amostragem realizado, sendo esperadas flutuações estatísticas relacionadas às diferentes amostras possíveis com as quais o classificador pode ter sido produzido e avaliado.

Definido um índice de desempenho, é desejável estimar com base na amostra

qual seria seu valor na população, a qual é desconhecida. Em outras palavras, produzido um classificador com base num conjunto de projeto particular (amostra), qual seria seu desempenho num cenário real (população). Sejam y e $\hat{y}$ os vetores real e estimado do índice, o primeiro obtido sobre a população, e o último estimado através da amostra disponível. Dois parâmetros utilizados para qualificar uma estimativa são a tendência e o seu erro quadrátivo médio (MSE - *mean square error*). Define-se a tendência de um estimador ($b[\hat{y}]$) como:

$$b[\hat{y}] = E[\hat{y}] - y, \tag{4.13}$$

ou seja, pela diferença entre o valor esperado do estimador, o qual é derivado com base em amostras da população, e o seu valor real (y). Para estimadores não-tendenciosos, $E[\hat{y}] = y$, logo o valor real do estimador pode ser inferido através das amostras. Em caso contrário, o estimador é considerado otimista ou pessimista [132], caso $b[\hat{y}] > 0$ ou $b[\hat{y}] < 0$, respectivamente.

O valor MSE de um estimador é definido como [52]:

$$MSE = E[(\hat{y} - y)^2], \tag{4.14}$$

ou seja, o MSE quantifica a flutuação do estimador com relação ao valor real, estando relacionado às diferentes amostras possíveis de uma mesma população com as quais $\hat{y}$ pode ser produzido. Note que a Equação 4.14 pode ser escrita na forma [141]:

$$MSE = E[(\hat{y} - y)^2] = (E[\hat{y} - y])^2 + E[(\hat{y} - E[\hat{y}])^2] = v[\hat{y}]^2 + \sigma_{\hat{y}}^2, \tag{4.15}$$

a qual foi decomposta em duas parcelas: uma relacionada à tendência ($v[\hat{y}]$); e outra à variância do estimador ($\sigma_{\hat{y}}^2$). Novamente, caso o estimador seja não-tendencioso, $MSE = \sigma_{\hat{y}}^2$, e o valor MSE pode ser estimado através das amostras. Em caso contrário, a variância ($\sigma_{\hat{y}}^2$) fornece um limite inferior para o erro quadrático médio (MSE) do estimador.

Outra forma de qualificar um estimador é através do intervalo de confiança [141]. Cada intervalo de confiança possui dois extremos, referidos como limites de confiança, e possui uma probabilidade associada, que é referida como nível de confiança (α), onde $0 < \alpha < 1$. Definidos um valor arbitrário para α e um critério para o cálculo dos limites de confiança, e supondo o estimador não-tendencioso, é possível garantir que $100(1 - \alpha)\%$ das amostras da população resultam em estimativas que

definem um intervalo de confiança que contém o valor real do parâmetro. Assim, dada uma amostra e calculado um intervalo de confiança, é provável, porém não garantido, que o valor real esteja no interior deste intervalo para $100(1-\alpha)\%$ das amostras possíveis. Caso o índice de desempenho possa ser associado à distribuição normal, é possível calcular os limites de confiança através de [141]:

$$\hat{y} \pm t_{\frac{\alpha}{2}[L]}\sigma_{\hat{y}}, \tag{4.16}$$

onde t é o valor crítico da distribuição t-*Student's* com L graus de liberdade.

A tendência observada na avaliação de desempenho de um classificador neural por um estimador arbitrário está fortemente relacionada ao seu aprendizado. Um pior aprendizado costuma resultar em maiores valores para a tendência. Conforme seção 4.1, para classificadores neurais, alguns fatores que impactam o aprendizado são: o tamanho e o conteúdo estatístico da amostra utilizada para o treinamento, a complexidade do classificador, o algoritmo de treinamento, e o critério de parada, quando aplicável [142].

A variância observada na avaliação de desempenho possui duas causas principais: a escolha da amostra (conjunto de treinamento) e os valores iniciais dos parâmetros considerados no treinamento do classificador [142]. Diferentes amostras de treinamento e valores de inicialização resultam, usualmente, em classificadores de desempenho distinto, em especial, na existência de restrições estatísticas. Caso o desempenho seja avaliado por validação cruzada somam-se a estas flutuações variações relacionadas à escolha da amostra de avaliação. Neste caso, uma forma possível de reduzir a tendência é aumentar o número de eventos do conjunto de projeto, o que resulta numa melhora do aprendizado, porém reduz o número de eventos disponível para o conjunto de avaliação, e ocasiona um aumento da variância [116]. Este problema é referido como dilema viés-variância, sendo necessário estabelecer um compromisso entre estes interesses conflitantes [118].

Uma primeira forma de avaliar um classificador segundo a validação cruzada é dispor de um conjunto de projeto e avaliação representativos do problema. Se for suposto que não há tendência, o que é razoável em classificadores baseados num conjunto de projeto que possua uma caraterização estatística apropriada do problema e tenham sido devidamente dimensionados e treinados, tem-se $E[\hat{y}_i] = y_i$. Como o conjunto de avaliação é também suposto representativo, uma aproximação razoável

é $E[\hat{y}_i] \approx \hat{y}_i = y_i$, ou seja, o desempenho real pode ser estimado com base numa única amostra de avaliação. Neste caso, para a estimativa do valor MSE, utiliza-se um modelo estatístico apropriado para as saídas da rede, como por exemplo, o método delta [141] ou o estimador *sandwich* [143], quando o índice de desempenho é o erro numérico da saída da rede, ou modelos baseados em distribuição binomial [1], confome discutido na Seção 4.2.1, quando o desempenho é aferido através de erros de classificação.

Outra forma de estimar o desempenho de um classificador é através das técnicas de reamostragem. Através destas técnicas, o desempenho do modelo é previsto através da produção e avaliação de vários classificadores. Cada classificador possui um conjunto de projeto e avaliação particular, os quais são formados através de uma divisão arbitrária dos dados disponíveis. Neste caso, a estimação do desempenho para um total de L partições considera:

$$E[\hat{y}] \approx \bar{y} = \frac{1}{L}\sum_{i=1}^{L}\hat{y}_i, \tag{4.17}$$

ou seja, que o valor esperado do estimador pode ser aproximado por uma média dos valores obtidos para cada partição ($\hat{y}_i$). Outro parâmetro a ser considerado é a variância dos valores estimados, determinada através de:

$$\sigma^2_{\hat{y}} \approx \hat{\sigma}^2_{\hat{y}} = \frac{1}{L-1}\sum_{i=1}^{L}(\hat{y}_i - \bar{y})^2 \tag{4.18}$$

Caso o estimador $\hat{y}$ seja não-tendencioso, $y = E[\hat{y}] \approx \bar{y}$ e $MSE = \sigma^2_{\hat{y}} \approx \hat{\sigma}^2_{\hat{y}}$, que podem ser determinadas pela avaliação do classificador nas diferentes partições. As técnicas de reamostragem são atrativas para a avaliação de desempenho pois não presumem nenhum modelo estatístico para o classificador [124], não exigem o conhecimento a priori de conjuntos de projeto e avaliação que sejam representativos do problema e levam em consideração as flutuações inerentes ao conjunto de dados [118] e devido à inicialização do classificador neural [142].

Entre as técnicas de reamostragem, a diferença principal reside na forma de produção dos conjuntos de projeto e avaliação. Duas linhas principais podem ser identificadas: uma baseada em amostragem sem reposição; e outra, em amostragem com reposição. Entre representantes da primeira linha, tem-se a subamostragem aleatória [144], a técnica *k-fold* [145] e o método deixe-um-de-fora (*Leave-one-out* - *LOO*) [129, 146]. Para a última linha, tem-se a técnica *Bootstrap* [147].

Pela técnica de subamostragem aleatória, para n eventos, em cada uma das L partições, formam-se conjuntos de projeto e teste com p e $(n-p)$ eventos, respectivamente. Em geral, utiliza-se $p = \frac{2}{3}n$ ou $p = \frac{1}{2}n$. Como metade (ou um terço) dos eventos não é utilizada no projeto dos classificadores, o estimador *holdout* é pessimista [144]. Pela técnica *k-fold*, dos n-eventos, formam-se k grupos com $\frac{n}{k}$ eventos. Cada partição é constituída considerando $(k-1)$ grupos formando o conjunto de projeto; e o grupo restante, o conjunto de avaliação. Testes experimentais em [144] mostram que a tendência de estimadores derivados através desta técnica se reduzem com o aumento de k, afirmando que para k em torno de 10 a 20, tem-se estimadores com tendência desconsiderável para boa parte dos problemas práticos. Quanto à técnica do deixe-um-fora (LOO) [129, 146], para n dados, são formadas n partições, onde cada partição utiliza, dos n eventos disponíveis, $(n-1)$ eventos no conjunto de projeto e apenas 1 evento no conjunto de avaliação. A técnica LOO produz estimadores com baixa-tendência, porém de alta-variância, e possui custo computacional elevado, por envolver a obtenção de n modelos. Por fim, pela técnica de *Bootstrap*, diferentemente das propostas anteriores, os conjuntos de projeto e avaliação podem possuir eventos em comum, uma vez que a amostragem é realizada com reposição. Esta técnica é freqüentemente recomendada para problemas com pequenos número de eventos [148]. Para a derivação do valor do estimador pela técnica de *Bootstrap*, algumas propostas são: o *Bootstrap 0.632*, por Efron (1983) [149], e o *Bootstrap 0.632+*, por Efron e Tibshirani (1997) [150], que combinam otimismo com pessimismo, dado que o valor do estimador é obtido por uma composição dos valores obtidos para as partições de projeto e avaliação. Em problemas práticos, recomenda-se em torno de 50 a 1000 partições [151, 152], o que é crítico para aplicações com grande número de eventos, tais como o problema do sonar passivo.

Na prática, a qualificação de desempenho é realizada através dos valores de $\bar{y}$ e $\hat{\sigma}_{\bar{y}}$. Assim, supondo válidas as aproximações propostas pelas Equações 4.17 e 4.18, o efeito duma estimação tendenciosa é derivar valores inferiores ao real, caso o estimador seja pessimista; ou superiores, caso otimista. Para ambos os casos, pela Equação 4.18, a variância observada será sempre inferior a do estimador, o que resulta em intervalos de confiança subdimensionados [153]. Ainda que a tendência não seja desejável, todas as técnicas de amostragem anteriormente descritas produzem

estimadores tendenciosos. Trabalhos recentes discutem a inexistência de estimadores não-tendenciosos para a técnica de subamostragem aleatória [153] e por *k-fold* [154]. Estimadores alternativos, com menor tendência e estimativas mais acuradas para o valor do MSE, são propostos em [153, 155], porém não serão considerados neste trabalho, em razão de possuirem um custo computacional proibitivo para o problema de sonar.

4.3 Caracterização estatística do conjunto de dados

Realizar a caracterização estatística de um conjunto de dados visa, em geral, atender aos resquisitos de um problema particular, o qual orienta o processo de busca e análise da informação existente nos dados.

No problema da classificação de contatos, é desejável identificar, classe-a-classe, a existência de informação discriminante dos dados em diferentes classes. Como cada classificador explora na sua tomada de decisão um conjunto particular destas características, em geral, esta caracterização é baseada na análise de desempenho do próprio classificador a ser produzido com base nos dados.

Na caracterização do volume de informação discriminante será utilizado o próprio classificador como figura de mérito. Para esta análise buscar-se-á prever qual é o desempenho esperado de um classificador produzido com base nos dados disponíveis, assim como prover as incertezas de classificação a ele associadas. A determinação destas incertezas permite qualificar a classificação provida pelo classificador, inferindo quais flutuações de desempenho são esperadas para os diferentes cenários possíveis de operação. Característica desejável é que esta flutuação seja a menor possível, traduzindo uma característica de robustez do classificador, ou seja, que seu desempenho não se degrade, substancialmente, de acordo com o cenário de operação considerado.

Para indicar o conteúdo estatístico das classes será utilizada a eficiência de generalização, visto que um desempenho inferior quanto à generalização é esperado para as classes que apresentarem as maiores restrições estatísticas. Através das confusões cometidas pelo classificador, isto é, pela quantidade de eventos erronea-

mente classificados para cada classe, serão identificadas quais classes possuem uma classificação mais difícil. A identificação das restrições quanto à caracterização estatística das classes pode orientar novas aquisições, ou mesmo na estruturação do sistema de classificação, contribuindo para a escolha de parâmetros relativos ao pré-processamento, à rede neural classificadora e ao treinamento, tais como o número de neurônios, critério de parada, tipo de gradiente, entre outros.

A formação dos conjuntos de projeto e avaliação, neste trabalho, conforme descrito na seção 4.1, explora a técnica de validação cruzada. A produção destes conjuntos pode, no entanto, seguir diferentes critérios, o que resulta, para o conjunto de sonar, na produção e avaliação do classificador considerando diferentes enfoques, conforme discutido na Seção 2.1.3.

Para ambas modalidades, conforme Seção 4.2.2, inferir o desempenho do classificador exige partições dos dados com conteúdo estatístico expressivo, as quais não estão tipicamente disponíveis. Assim, para a estimação do desempenho, foi considerada a utilização de técnicas de reamostragem. Dentre as opções discutidas na seção 4.2.2, a técnica de subamostragem aleatória é aquela que produz conjuntos de avaliação com maior número de eventos, o que resulta numa ênfase à avaliação da generalização e menor variância ou flutuação estatística do estimador de desempenho, fatores que motivaram sua adoção para o problema do sonar.

Esta técnica é ainda pessimista, ou seja, as estimativas produzidas tendem a ser inferiores ao valor real, o que é atrativo numa aplicação militar por questões de segurança. Com respeito ao quantitativo de eventos por conjunto, os dados foram particionados, meio-a-meio, entre os conjuntos de projeto e avaliação.

Na estimação por subamostragem aletória, diferentes conjuntos são produzidos e, para cada um, um classificador é produzido e avaliado. Dentre os vários conjuntos disponibilizados, um deles está associado ao classificador de melhor desempenho, logo representa o melhor equilíbrio entre aprendizado e avaliação, provendo o par que, dentre os pares considerados, melhor representa o conjunto de dados. Assim, a subamostragem aletória pode ser utilizada para a identificação de conjuntos representativos dos dados, a despeito da eficácia deste procedimento ser dependente, fortemente, do número de ensaios realizados.

Conforme discutido na Seção 4.1, usualmente, o desempenho de classifica-

dores neurais é sensível aos valores iniciais dos parâmetros. Assim, na flutuação estatística avaliada por critérios de reamostragem, há uma contribuição das inicializações. Para a avaliação estatística é desejado, no entanto, quantificar as variações de desempenho com respeito ao conteúdo dos dados e não aquelas devidas ao treinamento. Assim, visando reduzir a contribuição da inicialização nesta avaliação, para cada arquitetura de rede estudada foram treinados 5 classificadores, cada qual com uma inicialização distinta. Deste conjunto foi selecionado, através de índice específico, o classificador de melhor desempenho, o qual foi considerado na estimação de desempenho por reamostragem.

Por este critério, para a identificação da melhor inicialização, é necessário definir um índice de desempenho. A seleção deste índice deve considerar, para o conjunto de sonar, a inexistência de maiores informações sobre os dados, tais como a probabilidade de ocorrência e a importância relativa de cada classe. É interessante utilizar um índice que possua baixo custo computacional e reflita um equilíbrio entre o desempenho das diferentes classes. Dentre os índices discutidos na seção 4.2.1, a eficiência SP foi selecionada por melhor refletir este conjunto de características.

4.4 Resultados de classificação

Para a análise estatística dos dados foram considerados classificadores de 10 a 40 neurônios na camada intermediária utilizando os dados pré-processados segundo cadeia descrita na seção 2.1.2. Esta faixa mostrou, de acordo com alguns ensaios realizados e trabalhos anteriores [5], um compromisso atrativo entre a estatística disponível nos dados, a complexidade e a eficiência de classificação das redes.

A avaliação dos classificadores, conforme discussão anterior, considerou a técnica de subamostragem aleatória. Na literatura não foram encontradas referências que discutissem critérios para a definição do número de partições. Usualmente, trabalhos que utilizam esta técnica consideram um número arbitrário de partições, não realizando uma discussão mais aprofundada das razões desta escolha. Para o problema do sonar, foram utilizadas 10 partições, número o qual acredita-se permitir uma avaliação satisfatória das eficiências de classificação, possuindo um custo computacional factível ao problema.

Na Figura 4.1 são apresentadas as eficiências e as incertezas, estimadas pelo valor médio e pelo valor RMS das eficiências SP obtidas para os 10 ensaios considerados pela técnica de subamostragem aleatória, respectivamente, para redes de 10 a 40 neurônios na camada intermediária. É possível perceber diferenças relevantes nas eficiências e incertezas dos classificadores baseados na seleção por espectros e corridas. Para a seleção por espectros, as eficiências SP situam-se entre $(92,1 \pm 0,5)\%$ e $(95,6 \pm 0,2)\%$; enquando na seleção por corridas, entre $(80,6 \pm 1,6)\%$ e $(83,1 \pm 0,9)\%$, para 10 e 40 neurônios, respectivamente. Quanto às incertezas das eficiências SP, os valores associados à seleção por corridas são de 3 a 5 vezes superiores que os verificados para a seleção baseada em espectros. O crescimento da curva associada à seleção por espectros é maior que da curva associada à seleção por corridas: a primeira, de $3,5$ e $1,0$; enquanto a segunda, de $2,1$ e $0,4$ pontos percentuais, considerando as redes de 10 e 25, e de 25 e 40 neurônios, respectivamente. Para ambas as curvas, há uma tendência na estabilização dos valores das eficiências, mais marcante para a seleção baseada em corridas, à medida que redes com um maior número de neurônios são consideradas, o que é coerente com resultados da literatura [19].

Para avaliar o desempenho classe-a-classe dos classificadores foram consideradas redes com 10, 25 e 40 neurônios, que correspondem aos extremos e meio da faixa considerada na análise anterior. Para uma seleção baseada em espectros, a Figura 4.2(a) exibe as eficiências e incertezas de classificação de cada classe. Na Figura 4.2(b) são destacadas as incertezas de classificação, enquanto a Tabela 4.1 resume os valores utilizados na elaboração destes gráficos. Para as redes com 10 e 25 neurônios, o pior desempenho é obtido pela classe C, com eficiências de $(85,2 \pm 1,6)\%$ e $(91,8 \pm 1,1)\%$, respectivamente, enquanto que, para a rede de 40 neurônios, tem-se a classe A $(92,6 \pm 1,0)\%$ com pior desempenho. A classe de melhor desempenho é a E, com eficiências entre $96,6$ e $98,3$, e uma incerteza de $\approx 0,5$ pontos percentuais para as três topologias. Para todas as topologias, as classes de identificação mais críticas foram a A, B, C e H, que estão associadas às menores eficiências. As incertezas de classificação se situaram entre $\approx 0,4$ (classe E para a rede com 40 neurônios) e $\approx 1,6$ (classe C para a rede com 10 neurônios).

As eficiências e incertezas obtidas para a seleção baseada em corridas são apresentadas na Figura 4.3 e na Tabela 4.1. De forma similar à seleção baseada em

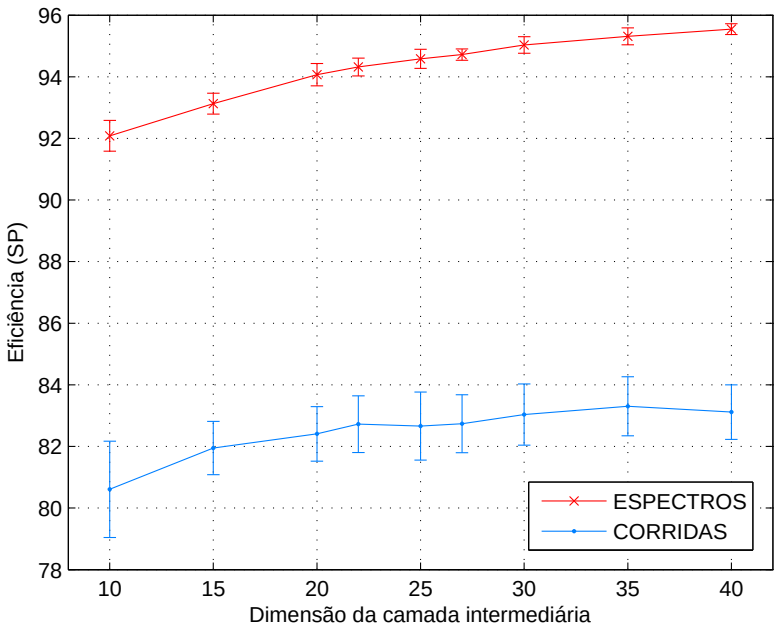
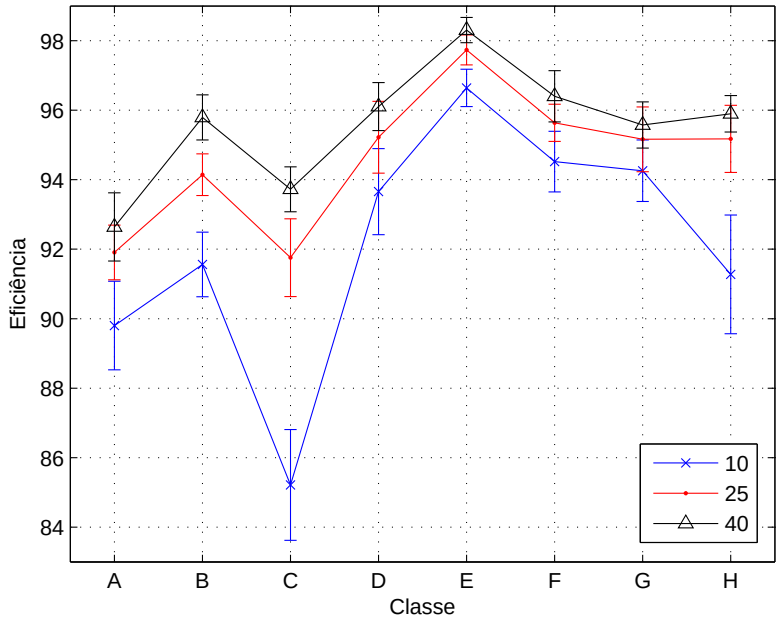
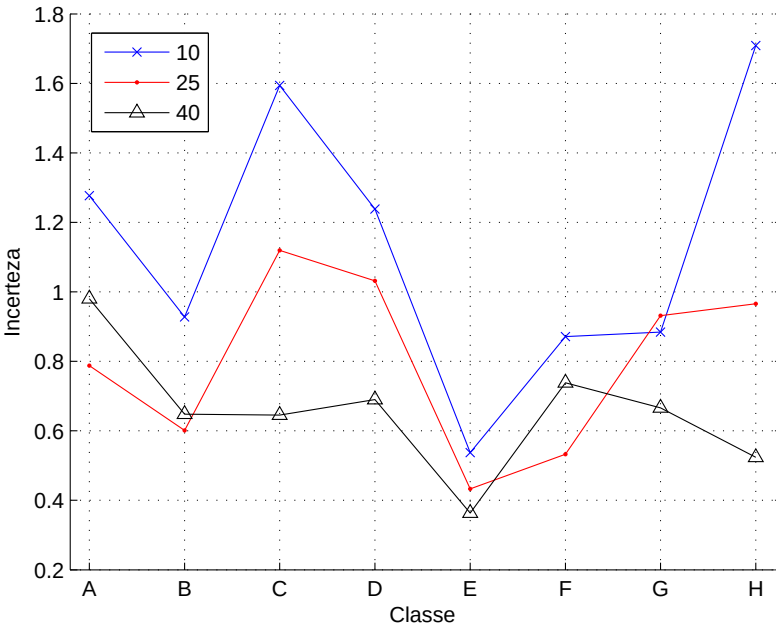


Figura 4.1: Eficiências de generalização (SP) do classificador estimadas por subamostragem aleatória, considerando a seleção por espectros e arquiteturas de rede com 10 a 40 neurônios na (única) camada intermediária.



(a)



(b)

Figura 4.2: Eficiências (a) e incertezas de classificação (b), classe-a-classe, para a seleção baseada em espectros, considerando redes com 10, 25 e 40 neurônios na única camada intermediária.

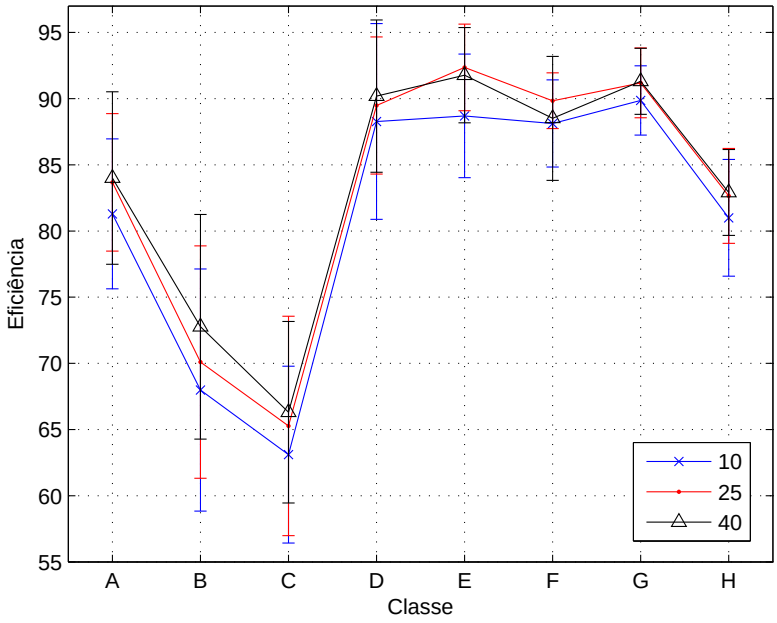
espectros, o pior desempenho é obtido para a classe C, nas três topologias consideradas, com eficiências de $(63,1 \pm 6,8)\%$ e $(66,3 \pm 6,9)\%$ para 10 e 40 neurônios, respectivamente. O melhor desempenho é obtido para a classe E, entre $88,7\%$ e $91,8\%$ pontos percentuais, enquanto as classes A, B, C e H apresentaram os menores valores de eficiência para todas topologias. Em relação às incertezas, os valores se situaram entre 2,1%, para a classe F e uma rede com 25 neurônios na camada escondida; e 9,1% para a classe B quando uma rede com 10 neurônios nesta mesma camada é utilizada.

Na comparação entre a seleção por espectros e corridas, para as três topologias consideradas, as maiores diferenças de desempenho foram verificadas para as classes B e C. Para a classe B, a seleção por espectros resulta em classficadores de 23 a 24 pontos percentuais mais eficientes que a seleção baseada em corridas. Em relação à classe C, estas diferenças situam-se entre 22,1 e 27,5. Diferenças menores são obtidas para as classes A e H: entre $9,8$ e $11,3$ para a classe A; e na faixa de $10,3$ a 13 para a H. Para a classe E, de melhor desempenho, as diferenças foram de $5,3$ a $7,9$.

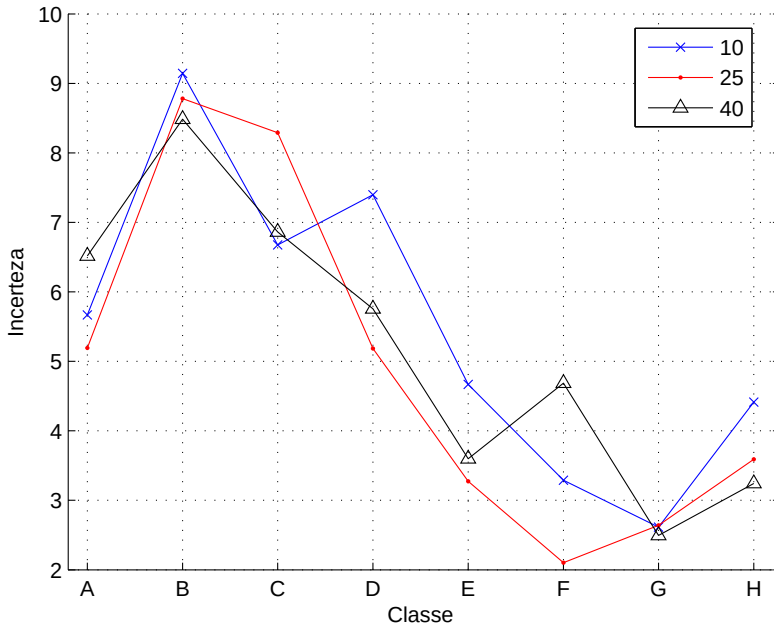
Tabela 4.1: Eficiências de classificação (%), classe-a-classe, para a seleção baseada em espectros e em corridas para redes com 10, 25 e 40 neurônios na única camada intermediária

Seleção baseada em espectros								
n	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
10 n	$89,8 \pm 1,3$	$91,6 \pm 0,9$	$85,2 \pm 1,6$	$93,7 \pm 1,2$	$96,6 \pm 0,5$	$94,5 \pm 0,9$	$94,2 \pm 0,9$	$91,3 \pm 1,7$
25 n	$91,9 \pm 0,8$	$94,1 \pm 0,6$	$91,8 \pm 1,1$	$95,2 \pm 1,0$	$97,7 \pm 0,4$	$95,6 \pm 0,5$	$95,6 \pm 0,9$	$95,2 \pm 1,0$
40 n	$92,6 \pm 1,0$	$95,8 \pm 0,6$	$93,8 \pm 0,6$	$96,1 \pm 0,7$	$98,3 \pm 0,4$	$96,4 \pm 0,7$	$95,6 \pm 0,7$	$95,9 \pm 0,5$
Seleção baseada em corridas								
n	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
10 n	$81,3 \pm 5,7$	$68,0 \pm 9,1$	$63,1 \pm 6,8$	$88,3 \pm 7,4$	$88,7 \pm 4,7$	$88,5 \pm 3,3$	$88,9 \pm 2,5$	$81,0 \pm 4,4$
25 n	$84,0 \pm 5,2$	$70,1 \pm 8,8$	$65,3 \pm 8,3$	$89,5 \pm 5,2$	$92,4 \pm 3,3$	$89,9 \pm 2,1$	$91,3 \pm 2,6$	$82,9 \pm 3,6$
40 n	$81,3 \pm 6,5$	$72,8 \pm 8,5$	$66,3 \pm 6,9$	$90,2 \pm 5,8$	$91,8 \pm 3,6$	$88,5 \pm 4,7$	$91,3 \pm 2,5$	$82,9 \pm 3,2$

A identificação das confusões cometidas pelo classificador pode ser um instrumento útil na identificação de quais classes possuem uma maior demanda quanto a novas aquisições, e fornecer, a priori, ao operador de sonar, uma informação a ser explorada, em cenários mais críticos, na tomada de decisão, em especial quando através de instrumentos adicionais o operador considerar que a classe identificada



(a)



(b)

Figura 4.3: Eficiências (a) e incertezas de classificação (b), classe-a-classe, para a seleção baseada em corridas considerando redes com 10, 25 e 40 neurônios na camada intermediária.

pelo sistema para o contato for duvidosa.

Num sistema de 8 classes, há 7 confusões possíveis por classe, porém, para maior simplicidade, os resultados apresentados contemplarão, apenas, a maior das confusões observadas para cada classe. Para quantificar estas confusões, foram considerados classificadores com 15, 25 e 40 neurônios na camada intermediária, produzidos pela partição que melhor representa os dados, identificada pelo processo descrito na seção 4.3.

Na Tabela 4.2, para uma classe arbitrária X, definida pela coluna, é indicado a qual classe são mais freqüentemente atribuídos (erroneamente) os eventos dela provenientes, juntamente com uma faixa de valores, que indica, em termos percentuais, a quantidade de eventos associados a estas confusões para as três topologias analisadas.

Tabela 4.2: Maiores confusões por classe de contato (em taxa de confusão percentual) para a seleção baseada em espectros e em corridas (Veja o texto).

Seleção baseada em espectros							
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
C(5,0 - 7,1)	G(1,2 - 2,7)	A(2,8 - 5,5)	C(1,6 - 2,5)	C(0,5 - 0,8)	C(0,8 - 2,6)	B(1,4 - 3,0)	C(1,3 - 1,9)
Seleção baseada em corridas							
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
C(22,2 - 35,4)	C(7,4 - 16)	H(6,3 - 8,2)	G(0,1 - 3,6)	C(2,8 - 13,7)	C(6,2 - 9,7)	B(3,8 - 6,9)	C(3,9 - 5,3)

É possível perceber que, para ambas seleções, as maiores confusões são verificadas para os eventos da classe A, que são classificados como oriundos da classe C. Fato similar ocorre para as classes E, F e H. Para a seleção por espectros, os eventos da classe B tendem a ser classificados como da classe G, e vice-versa. O mesmo ocorre na seleção por corridas com respeito às classes C e H. Novamente, as taxas de confusão associadas à seleção baseada em corridas mostraram-se maiores que a baseada em espectros.

Dos resultados anteriores, verifica-se que a seleção por corridas, na qual, os conjuntos de projeto e avaliação possuem diferentes condições operativas, provê um teste mais severo quanto à generalização, onde menores eficiências, maiores flutuações e uma maior dependência quanto à composição de ambos conjuntos são observadas. Através das diferenças de desempenho verificadas para as seleções por

espectros e corridas, tanto em termos dos valores médios quanto pelas incertezas, é possível identificar restrições estatísticas críticas com relação à caracterização das diferentes condições operativas das classes. Desta forma, no projeto de um sistema de classificação que utilize a base de dados em análise, os conjuntos de projeto e avaliação devem ser criteriosamente selecionados.

Para ambas modalidades de seleção, foram identificadas as mesmas classes como críticas, e as confusões cometidas mostraram-se bastante similares. Através da seleção por corridas, verifica-se que a classe mais crítica é a C [2], a qual é seguida pela B, A e H. Significativa confusão é verificada entre as classes A e C, enquanto os pares de classes $\{C, H\}$ e $\{B, G\}$ apresentaram confusões consideráveis. Estes resultados sinalizam que novas aquisições devam considerar um maior número de condições operativas para cada navio, em especial para as classes mais críticas tais como a A e a C.

[2]Dado que nenhuma informação sobre as características das classes foram disponibilizadas pelo IPqM, supõe-se que o pior desempenho da classe C possa estar associado uma maior heterogeneidade dos navios nela envolvidos. Cabe ainda observar que a classe C possui o maior número de navios entre as demais (cinco, no total).

Capítulo 5

Seleção dos conjuntos de projeto e avaliação

Conforme discutido anteriormente, para a produção do sistema de classificação é necessária a partição dos dados em conjuntos de projeto e avaliação, cujo conteúdo está relacionado diretamente com o aprendizado e a avaliação do sistema. No problema do sonar, através dos resultados de classificação obtidos no Capítulo 4, restrições estatísticas quanto à caracterização das classes foram identificadas, em especial para a seleção por corridas, exigindo uma criteriosa seleção destes conjuntos, sob o prejuízo de comprometer o aprendizado ou prover uma estimativa não-realística do seu desempenho. Nesta escolha, um compromisso apropriado entre as qualidades de aprendizado e de avaliação mostrou-se necessário.

Um critério intuitivo para a seleção dos conjuntos é por sorteio, conforme apresentado na seção 4.3. Trata-se, no entanto, de um processo fortemente baseado em tentativa e erro, cujo desempenho é dependente do número de ensaios realizados, e o custo computacional é proibitivo, visto que cada sorteio demanda o treinamento de um classificador.

Uma das propostas deste trabalho é realizar a seleção por espectros e por corridas através de agrupamentos (*clustering*). As técnicas de agrupamento particionam os dados em grupos de características similares, aferidas por funções matemáticas específicas [156]. Assim, de posse dos grupos identificados pelo agrupamento, a partição dos dados pode resultar em conjuntos de projeto e avaliação que melhor retêm a estatística existente nos dados sob análise [1, 156].

Para a produção de agrupamentos, há um número significativo de algoritmos na literatura, cada qual explorando uma estratégia particular. De acordo com a estratégia adotada e os parâmetros envolvidos, diferentes características e grupos podem ser identificados, o que impacta na seleção dos conjuntos. Para o problema de sonar, que é crítico, a escolha do algoritmo e a seleção de seus parâmetros deve, portanto, ser cuidadosa.

Na seleção por espectros (vide seção 2.1.3), a formação dos conjuntos pode ser derivada, diretamente, com base na pertinência dos eventos aos grupos definidos pelo agrupamento. Para a seleção por corridas, no entanto, é necessário formar agrupamentos de corridas, que representam subconjuntos dos dados. Neste caso, é necessário dispor de critérios para a geração destes agrupamentos, assim como índices que identifiquem corridas similares. Como não foram encontradas referências tratando este problema, foi desenvolvido um critério que, com base no agrupamento produzido para a seleção por espectros, produz um segundo agrupamento, onde as corridas, através de vetores representativos, são agrupadas por similaridade.

Este capítulo é iniciado com uma discussão da aplicação de técnicas de agrupamento à seleção por espectros. Nesta abordagem, características gerais dos agrupamentos são discutidas, em especial com respeito às técnicas seqüencial e hierárquica, sendo propostos e avaliados critérios para a seleção dos parâmetros envolvidos. Em seguida são discutidos e avaliados os critérios propostos para a seleção por corridas.

5.1 Análise de agrupamentos para a seleção por espectros

Para o problema de sonar, como o número de eventos disponível é significativo, optou-se por algoritmos de produção de agrupamentos simples e eficazes, com custo computacional compatível com a aplicação, preferencialmente não-paramétricos, em virtude das restrições estatísticas que se anunciam. Foram selecionadas as técnicas de agrupamento seqüencial [1] e hierárquico [1] que, por possuirem estratégias significativamente diferentes, podem evidenciar características distintas dos dados.

Usualmente, em espaços de dados de dimensão reduzida, as técnicas de agrupamento realizam uma melhor identificação de grupos de características estatísticas similares existentes nos dados. No problema do sonar, uma primeira extração de informação relevante ao problema é realizada pelo sistema de pré-processamento, descrito na seção 2.1.2, que resulta em janelas espectrais com 557 componentes. Como a dimensão destas janelas é ainda elevada para a produção dos agrupamentos, foi necessário realizar sua compactação, o que pressupõe a escolha de uma técnica e da quantidade de informação que será preservada nos dados compactados. Uma técnica de compactação consagrada é a análise das componentes principais [157], que enfatiza a representação dos dados. Uma das propostas deste trabalho é avaliar a compactação baseada em componentes discriminação [158], a qual privilegia informação discriminante entre as classes, e tende a produzir níveis de compactação mais elevados [5]. Para ambas análises, faz-se necessária a definição do número de componentes utilizadas na compactação, parâmetro que está relacionado à quantidade de informação retida nos dados compactados.

Os parâmetros utilizados nos algoritmos de produção de agrupamentos possuem relação direta com o formato dos agrupamentos produzidos, logo sua escolha é bastante relevante ao processo de seleção dos conjuntos. No agrupamento seqüencial, a escolha do raio de cada grupo influencia no número de grupos criados. Tanto menor o valor do raio, maior é o número de grupos, visto que um maior nível de detalhe, logo uma granularidade mais fina é considerada para a formação dos grupos. Usualmente, a escolha do raio é realizada de forma heurística, com base em alguns ensaios, de acordo com as especificidades do problema. Para o agrupamento hierárquico, tem-se uma estrutura em níveis, referida como dendrograma [1], onde cada nível define um agrupamento de dada granularidade. Neste caso, cabe ao especialista identificar qual granularidade melhor retrata a estrutura existente nos dados.

Em ambos os casos, uma granularidade grosseira ou demasiado fina resulta em agrupamentos deficientes e não realísticos. Dois casos extremos desta categoria são: todos os dados num mesmo grupo ou um grupo para cada dado. Em conjuntos com restrições quanto à sua caracterização estatística, o impacto de um agrupamento deficiente no processo de seleção é maior, dado que a existência de cenários raros ou críticos é mais provável e tais cenários podem não ser evidenciados pelo agrupamento

produzido.

A seleção da granularidade dos agrupamentos não é um problema trivial, com solução fechada na literatura. Freqüentemente, esta seleção é realizada pela inspeção do especialista aos grupos formados, a qual é baseada num conhecimento a priori dos dados. Para conjuntos de dados com um número expressivo de eventos e pouca informação sobre suas características, tais como o conjunto em estudo, esta seleção pode ser crítica, ou mesmo, inviável.

Na literatura, a seleção da granularidade é, normalmente, baseada na identificação do número de grupos existentes nos dados. Freqüentemente, utilizam-se critérios relativos [1], ou seja, são produzidos vários agrupamentos e identificado, através de heurísticas e da comparação de valores de índices específicos produzidos sobre os agrupamentos, qual deles melhor reflete a estrutura existente nos dados. Na literatura, há diferentes propostas de índices [159, 160, 161, 162], o que torna a seleção do agrupamento por índices dependente das características do índice escolhido.

Na seleção dos conjuntos de projeto e avaliação baseada em agrupamento, não há garantia de que o agrupamento identificado por um dado índice seja o que resulta num melhor desempenho do classificador, em especial para conjuntos de dados complexos, tal qual o de sonar. Perguntas possíveis seriam: qual índice é o mais adequado a esta seleção? Qual critério utilizar para a produção dos diferentes agrupamentos necessários?

Estes fatores motivaram a elaboração de dois critérios: um aplicável ao agrupamento seqüencial e outro ao hierárquico. Em ambos critérios, o desempenho do classificador é a figura de mérito que orienta a escolha dos parâmetros. Aspecto interessante é que esta proposta já inclui o efeito de como os grupos do agrupamento são utilizados para a formação dos conjuntos, assim como provê qual é o desempenho do classificador para o conjunto selecionado pelo critério. Estes critérios serão discutidos nas seções referentes à descrição das técnicas seqüencial e hierárquica.

5.1.1 Considerações gerais sobre agrupamentos

A idéia básica envolvida nas técnicas de agrupamento é revelar a organização intrínseca existente nos dados, o que permite estabelecer relações de semelhança

e diferença entre os padrões, assim como a derivação de conclusões úteis sobre os dados.

O agrupamento consiste numa das mais primitivas atividades humanas, necessária em razão da enorme quantidade de informação recebida diariamente, cujo processamento individual seria impossível. Assim, a mente humana caracteriza entidades em agrupamentos, onde cada agrupamento é caracterizado pelos atributos comuns partilhados pelas diferentes entidades [1].

Um primeiro passo para a produção de um agrupamento consiste em realizar uma correta representação dos padrões, isto é, das diferentes informações trazidas pelos dados [156]. Passo posterior é a seleção de características ou variáveis a serem consideradas na formação dos agrupamentos. Nesta etapa, realiza-se a redução do número de variáveis envolvidas, que pode se basear na seleção das variáveis mais relevantes à caracterização do problema ou na extração de características, a última realizada pela aplicação de transformações de compactação dos dados. Em ambos os casos, busca-se eliminar a informação redundante ou descorrelacionada com a solução do problema de interesse, dado que as técnicas de agrupamento costumam apresentar melhor desempenho para espaços de dados de dimensão reduzida [156].

Na formação dos agrupamentos, faz-se necessário aferir relações de similaridade entre dados e grupos. Freqüentemente, para variáveis reais, esta similaridade é medida com base em distância geométrica. Há, portanto, a crença de que as similaridades existentes entre as características dos dados sejam apropriadamente refletidas no espaço dos vetores de características. Em outras palavras, tanto mais similares dois eventos, mais próximos geometricamente seriam seus vetores de características. A avaliação de similaridade exige, portanto, medir a distância entre dois vetores. Um critério de distância bastante geral deve-se a Minkowski [1], o qual, para dois vetores $\mathbf{x}$ e $\mathbf{y}$ de dimensão l, é definido como:

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{l} |x_i - y_i|^p \right)^{\frac{1}{p}}, \tag{5.1}$$

onde x_i e y_i são componentes dos vetores $\mathbf{x}$ e $\mathbf{y}$, respectivamente, e p é um número inteiro qualquer, igual ou maior que 1. Um caso particular, usualmente utilizado, é a distância euclidiana, que é obtida para $p = 2$. Há ainda critérios que consideram pesos diferentes para cada par de componentes, entre eles: a distância euclidiana ponderada [163] e a distância de Mahalanobis [49, 156].

Definida uma medida de similaridade, cabe a seleção de um algoritmo para a produção do agrupamento. Diferentes algoritmos estão disponíveis na literatura, cada qual explorando uma estratégia particular. Adicionalmente, número significativo de algoritmos exigem a definição de parâmetros que possuem uma relação direta com a forma dos agrupamentos produzidos. Cabe proceder uma seleção cuidadosa destes parâmetros, a fim de que o agrupamento produzido reflita a real estrutura dos dados.

Por fim, dado que as técnicas de agrupamento propõem a existência de relações entre os dados, as quais podem ou não existir, cabe sua verificação, procedimento que é conhecido como validação. Após a validação, o especialista deriva, com base no agrupamento final, suas conclusões sobre os dados.

5.1.1.1 Extração de características

Para um grande número de aplicações em variadas áreas, faz-se necessário lidar com conjuntos de dados de dimensão elevada. Dimensão elevada não é, normalmente, sinônimo de quantidade ou qualidade de informação. Muitas aplicações apresentam um grande número de variáveis redundantes, isto é, que tendem a carregar informações semelhantes. Freqüentemente, apenas um pequeno subconjunto de informações ou variáveis é relevante à aplicação. A extração de características visa selecionar esta informação, eliminando informações complementares que não contribuem ou mesmo podem atrapalhar a solução do problema.

No apêndice B, discute-se a extração linear de cararacterísticas, que pode ser representada por:

$$\mathbf{f} = \mathbf{T}^T\mathbf{x}, \tag{5.2}$$

onde $\mathbf{T}$ é uma matriz de dimensões k x n, referida como matriz de extração de características ou de compactação dos dados, pois tipicamente reduz a dimensão dos eventos de n para k, onde k corresponde ao número (reduzido) de características selecionadas.

Para a escolha das direções associadas às colunas de $\mathbf{T}$, usualmente, utiliza-se a análise de componentes principais (PCA) [157]. Segundo esta análise, direções privilegiadas com respeito à representação dos dados são extraídas com base na

matriz de correlação dos dados. Estas direções serão aqui referidas como direções de representação. A análise PCA é uma técnica extensivamente aplicada em variadas áreas, entre elas: detecção, estimação, reconhecimento de padrões, processamento de aúdio e vídeo, assim como na compactação de espaços de alta-dimensionalidade [63].

Dado que a produção dos agrupamentos tem como objetivo a classificação, as componentes principais podem não ser as mais apropriadas para a representação das diferenças existentes entre as classes, que são as características de interesse ao problema. Direções privilegiadas com respeito à classificação podem ser obtidas através da análise de componentes principais de discriminação (PCD), cuja extração pode ser realizada pelo treinamento de um classificador MLP segundo um processo diferenciado. As análises PCA e PCD são discutidas em maiores detalhes no Apêndice B. Uma das propostas deste trabalho é avaliar a produção de agrupamentos com base em dados compactados por componentes de discriminação, uma vez que a análise PCD tende a produzir níveis de compactação mais elevados, o que pode resultar numa seleção dos conjuntos melhor sucedida.

5.1.2 Agrupamento seqüencial

A técnica seqüencial aqui considerada [1, 164] utiliza a distância euclidiana como medida de similaridade entre os padrões, produzindo agrupamentos formados por um conjuntos de hiperesferas, cujo raio, referido como raio de vigilância, é definido pelo usuário. Cada hiperesfera define um grupo, sendo caracterizada, individualmente, pelas coordenadas do seu centro e pelo valor do seu raio. Diferentes grupos são criados, de forma que todos os dados fiquem contidos no interior de um dado número de hiperesferas. Por este algoritmo, cada evento pertence a uma única hiperesfera, sendo o número de hiperesferas ou grupos necessários determinado automaticamente.

Na Figura 5.1, é ilustrado o formato final de um agrupamento seqüencial para um conjunto de dados arbitrário. É possível perceber que o formato do agrupamento é dependente da escolha dos raios de vigilância. Ainda que cada grupo possa possuir um valor de raio particular, por maior simplicidade e em virtude do desconhecimento das distribuições estatísticas dos dados, será considerado um mesmo valor de raio

de vigilância para todos os grupos a serem formados.

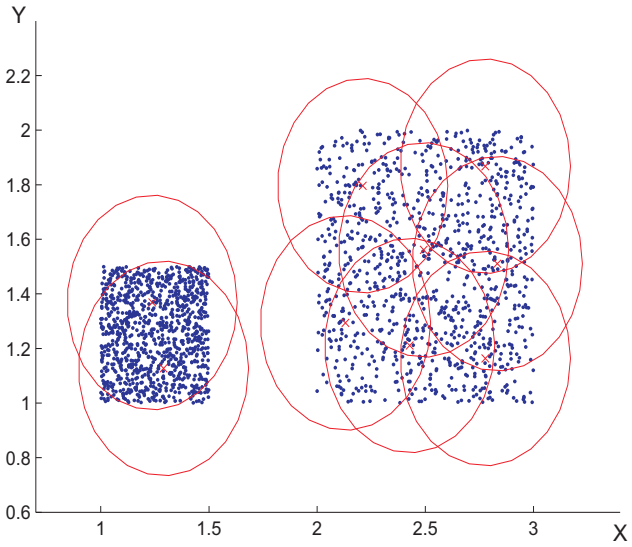


Figura 5.1: Agrupamento seqüencial sobre dados arbitrários

Para a produção do agrupamento seqüencial, os eventos são apresentados segundo uma seqüência aleatória, realizando-se uma forma de aprendizado competitivo, já que as hiperesferas ou grupos existentes disputam os dados. Para cada evento (espectro) apresentado, há duas alternativas: a criação de um novo grupo ou a incorporação do evento a um grupo existente. A alternativa a ser escolhida é definida pela avaliação da similaridade do evento aos grupos existentes. Para medida desta similaridade, adotou-se a função:

$$s_i = r_i^2 - ||\mathbf{x} - \mathbf{c}_i||^2, \tag{5.3}$$

onde r_i e $\mathbf{c}_i$ correspondem ao raio e ao centro do i-ésimo grupo existente. Se $s_i < 0$, o evento está no exterior da hiperesfera do i-ésimo grupo. Caso contrário, quanto mais próximo for o evento do centro da hiperesfera, maior será o valor de s_i. Caso o valor de s_i seja negativo para todos os grupos, o evento não se encontra no interior de nenhum grupo, criando-se um novo grupo, cujo centro é o próprio evento, com o raio de vigilância idêntico aos demais grupos. Havendo mais de um grupo com s_i positivo, o evento é incorporado ao grupo de maior valor de s_i, o qual é referido como vencedor. Nesta incorporação, o valor do centro do grupo vencedor é atualizado pela fórmula:

$$\mathbf{c}_i = \mathbf{c}_i + \eta||\mathbf{x} - \mathbf{c}_i||, \tag{5.4}$$

onde η é uma constante, referida como fator de aprendizado, escolhida de forma que: $0 < \eta < 1$. Através desta fórmula, o centro do grupo vencedor se aproxima do evento, de forma que, ao final do treinamento, as coordenadas do centro de cada grupo correspondam ao baricentro dos eventos a ele pertencentes.

Pelo algoritmo assim descrito, uma vez criado um grupo, ele será mantido no agrupamento, ainda que, em razão da evolução do aprendizado, parte ou todos seus eventos possam ter sido absorvidos por outros grupos. Este problema é facilmente observado em simulações com dados bidimensionais. Para ilustrar este problema, na Figura 5.2 apresenta-se um agrupamento seqüencial, onde são destacados dois grupos, cujos eventos também se encontram no interior de outros grupos do agrupamento final. Tratam-se, portanto, de grupos artificiais, cuja retirada do agrupamento não produziria nenhum prejuízo, uma vez que seus eventos seriam absorvidos pelos demais grupos restantes.

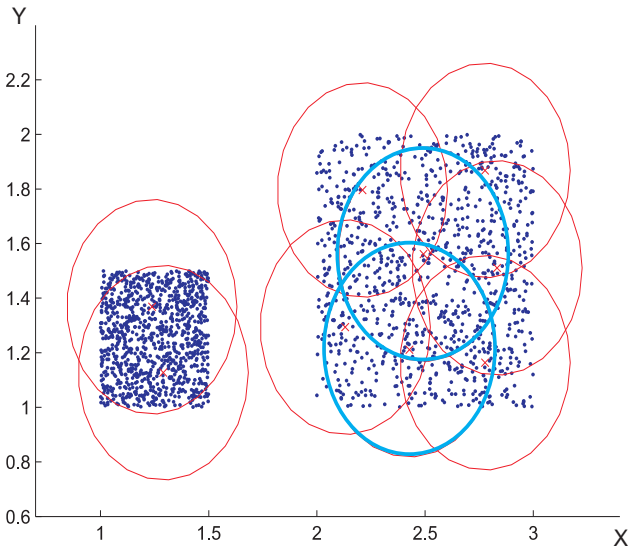


Figura 5.2: Identificação de grupos redundantes no agrupamento seqüencial de dados arbitrários

Faz-se necessário, portanto, um processo de identificação dos grupos redundantes, aqui referido como crítica, o qual pode ser realizado durante o processo de aprendizado ou após sua conclusão. Uma crítica simples é proposta em [2], a qual sugere que grupos com pequeno número de eventos sejam eliminados ao longo do processo de treinamento. Fica a pergunta: como determinar um valor apropriado

de eventos para esta eliminação? Ainda que este valor seja definido, como discernir entre grupos redundantes e aqueles que represententam características raras dos dados?

A criação de grupos redundantes pelo agrupamento seqüencial é um problema que não costuma ser abordado na literatura, motivando, assim, a proposição de uma crítica. Por esta crítica, através de um processo iterativo, é identificado o número mínimo de grupos necessário para encapsular todos os dados. Para esta identificação, são atribuídos níveis de relevância aos grupos, definidos pelo número de eventos que cada um possui. Inicialmente, é formado um conjunto de referência que contém todos os dados. Com base neste conjunto, o grupo mais relevante, isto é, aquele com maior número de eventos, é identificado. Em seguida, os eventos pertencentes a este grupo são eliminados do conjunto de referência, repetindo-se este processo, até que o conjunto esteja vazio. Os grupos não identificados por este processo são considerados redundantes e eliminados do agrupamento.

Caso um conjunto de grupos seja eliminado, provavelmente os centros dos grupos não correspondem mais ao baricentro dos eventos a ele pertencentes. Assim, faz-se necessário um ajuste dos seus centros, exceto quando a crítica é realizada durante o processo de obtenção do agrupamento [1]. Para a crítica pós-treinamento, o ajuste dos centros pode produzir modificações na estrutura do agrupamento, o que exige uma nova aplicação da crítica. Uma forma de identificar estas modificações é avaliar o número de grupos existentes, cuja alteração sinaliza que esta seqüência (crítica e ajuste) deve ser repetida. Como o ajuste dos centros pode produzir eventos sem grupos associados, limitar o número de eventos sem grupo é outro critério útil de parada.

5.1.2.1 Escolha do raio de vigilância

No agrupamento seqüencial, um parâmetro que impacta diretamente na formação dos grupos é o raio de vigilância. Geometricamente, tanto menor o valor do raio, maior é o número de grupos produzidos, já que mais hiperesferas são necessárias para encapsular os dados. Outra interpretação diz respeito ao nível de similaridade existente entre os dados que compartilham um mesmo grupo. Para grupos com raios

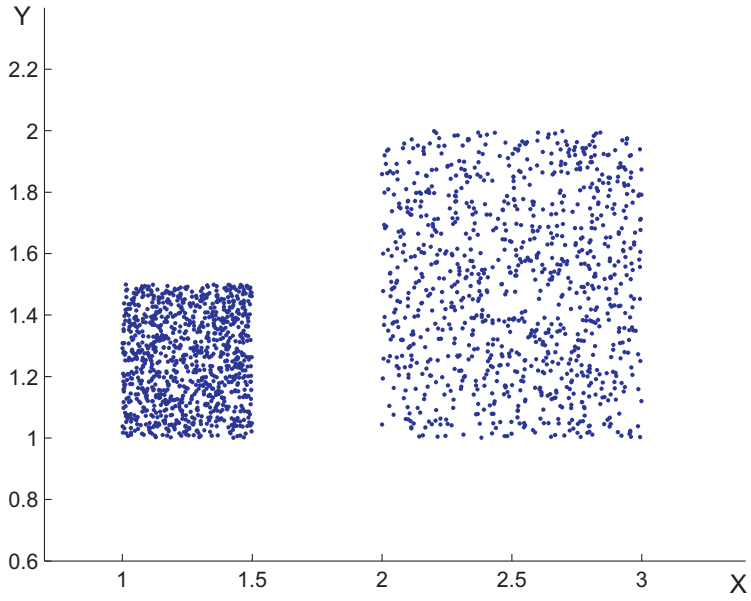
[1] Neste caso, o ajuste dos centros é realizado pelo próprio algoritmo de produção do agrupamento.

menores, há uma maior similaridade entre os eventos a eles pertencentes do que no caso de grupos obtidos a partir de raios de vigilância maiores. Assim, a escolha do raio de vigilância impacta na granularidade ou no nível de detalhe considerado para a construção do agrupamento. Para raios maiores, o algoritmo é menos detalhista, produzindo um menor número de grupos. Para raios menores, ocorre o contrário.

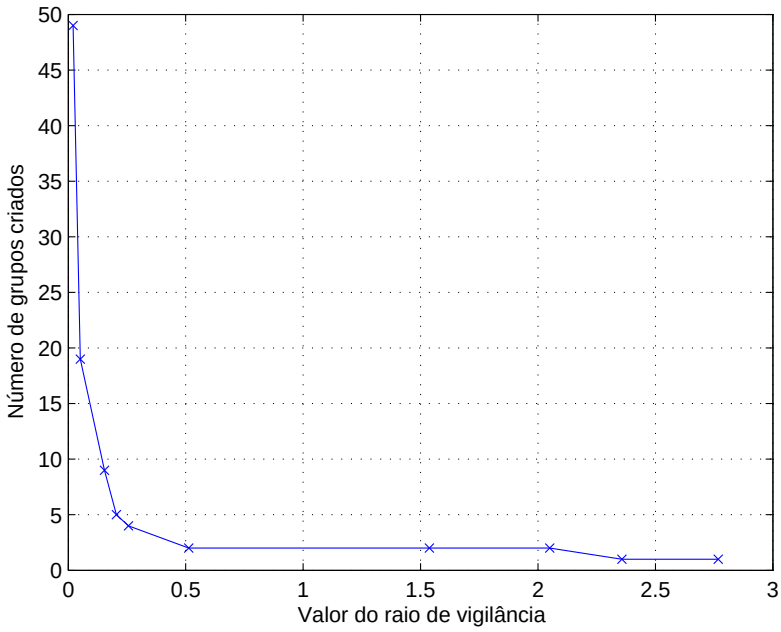
Como o agrupamento considerado destina-se à seleção dos conjuntos de projeto e avaliação, o valor do raio de vigilância deve ser cuidadosamente escolhido. A escolha de um valor de raio elevado pode originar grupos onde corridas ou navios notadamente distintos sejam tido como similares. Neste caso, é provável que cenários raros ou específicos não sejam evidenciados pelo agrupamento. Para valores de raios demasiado reduzidos, grupos irreais podem ser formados, não sendo identificadas, dada a crítica excessiva, similaridades relevantes entre os dados.

Definir um valor de raio ideal é uma tarefa complexa, já que pressupõe um conhecimento sobre os dados que, normalmente, não está disponível a priori. Em [1] é proposto um critério onde o valor do raio de vigilância é escolhido com base no número de grupos existentes no agrupamento. Para tal tarefa, define-se uma faixa de valores de raio, cujos extremos correspondem à distância mínima e máxima dos dados. Considerando valores nesta faixa, é elaborado um gráfico do número de grupos do agrupamento versus valor do raio de vigilância. O valor do raio a ser escolhido corresponde ao menor valor de raio pertencente a uma região plana do gráfico, o qual fornece uma estimativa do número de grupos existentes nos dados. Para fins de ilustração, na Figura 5.3 é exibida uma distribuição de dados arbitrária, assim com a curva do número de grupos criados versus valores de raio. Segundo o critério, o valor de raio a ser escolhido é 0,5, o qual produz um agrupamento com 2 grupos.

Ainda que atrativa por sua simplicidade, para um bom desempenho desta heurística é necessário que os diferentes grupos sejam compactos, bem separados e possuam dispersões de mesma ordem, o que é bastante improvável, em especial para dados provenientes de distribuições estatísticas complexas, como se apresentam os dados de sonar passivo. A identificação de uma região plana no gráfico, a qual pode não ser evidente, ou mesmo, não existir, é também um processo bastante subjetivo. Por fim, a escolha dos extremos da faixa de valores de raios baseada em valores



(a)



(b)

Figura 5.3: Ilustração do critério heurístico proposto em [1] para dados bidimensionais arbitrários: (a) dados e (b) curva do número de grupos criados versus raio de vigilância escolhido.

mínimos e/ou máximos de distâncias dos dados também não é indicada. Neste caso, a contra-indicação deve-se à susceptibilidade destas medidas extremas a eventos espúrios, comuns neste ambiente pelo alto-nível de ruído existente.

Uma estratégia alternativa é escolher os raios de vigilância com base na identificação do número de grupos existentes nos dados, cuja determinação é também um problema complexo. Na literatura, há diferentes estratégias e propostas para esta estimação [165]. Em linhas gerais, as diferentes propostas contemplam índices, descritos em maiores detalhes na seção 5.1.4, critérios baseados em comparações entre agrupamentos produzidos com base em diferentes partições dos dados, assim como algoritmos paramétricos, onde distribuições estatísticas, em geral, modelos gaussianos de mistura, são ajustadas sobre os dados.

Tibshirani (2000) [166] propõe a identificação do número de grupos existentes nos dados pela estatística GAP. Em linhas gerais, a estatística GAP realiza vários ensaios. Para cada ensaio, é suposto que o agrupamento possui um dado número de grupos, produzindo-se uma estatística baseada na comparação de vários agrupamentos de dados aleatórios e um agrupamento baseado nos dados originais. Através de uma análise subjetiva dos valores produzidos por esta estatística, define-se o número de grupos existentes. Na linha de algoritmos baseados em reamostragem, a estratégia consiste em comparar agrupamentos produzidos com base em diferentes partições dos dados, as quais são geradas por uma técnica de reamostragem arbitrária. Nesta comparação, através de diferentes propostas de estatística, o número de grupos existentes é identificado. Alguns algoritmos que exploram esta estratégia são o agrupamento consensual, de Monti et al.(2003) [167], e o *prediction strength*, de Tibshirani (2005) [168]. Seguindo a linha de modelagem estatística dos dados, Zhang (1990) [169] propõe a utilização de modelos gaussianos de mistura, para os quais a identificação do número de grupos existentes corresponde ao número de gaussianas utilizadas, definido através de índice específico.

Para problemas com grande número de eventos, uma primeira crítica às técnicas anteriores é o custo computacional, o qual está associado à produção de múltiplas partições, agrupamentos e modelos estatísticos sobre os dados. Para o último caso, freqüentemente, a estimação dos modelos não é um processo simples, em especial, quando há restrições estatísticas nos dados. Adicionalmente, a definição do raio de

vigilância através do número de grupos existentes, ainda que corretamente identificado, não implica que o agrupamento correspondente resulte numa seleção ótima dos conjuntos de projeto e avaliação do classificador. A escolha de um valor de raio adequado está muito mais relacionada à granularidade própria para o tratamento dos dados do que ao número de grupos existentes, em especial quando os dados possuem distribuições estatísticas complexas.

Estes fatores motivaram a proposição de um critério de seleção que utiliza o próprio classificador como figura de mérito. A idéia básica desta proposta é produzir diferentes agrupamentos, cada um baseado num valor de raio particular. Para cada agrupamento, é produzido um par de conjuntos de projeto e avaliação, e um classificador é treinado e avaliado. O classificador de maior eficiência de generalização identifica o melhor par de conjuntos, logo o agrupamento e raio mais adequados. Por este critério, a seleção do raio já considera como o agrupamento é explorado na formação dos conjuntos, e a granularidade mais apropriada para o tratamento dos dados, de forma independente ao número de grupos existentes ou da técnica utilizada para sua identificação.

Para definir quais valores de raio serão avaliados, explora-se a idéia de granularidade, a qual está relacionada, conforme discussão anterior, ao número de grupos existentes. A escolha dos raios pode ser realizada através da produção e avaliação de uma curva do número de grupos criados em função de diferentes valores do raio de vigilância. Assim, dentre os vários valores de raio cobertos por esta curva, é possível selecionar quais resultam em agrupamentos com granularidades distintas. Para o exemplo discutido na Figura 5.3, valores de raio na faixa de 0,5 a 2,0 resultam num agrupamento de mesma granularidade, já que estão associados a agrupamentos com 2 grupos. Para valores de raio de 0,2 e 0,5, os agrupamentos possuem uma granularidade similar, visto que possuem 5 e 4 grupos, respectivamente. Variações expressivas ocorrem para valores de raio inferiores a 0,2, como, por exemplo, para 0,15 e 0,05, que apresentam 9 e 19 grupos. Assim, a curva do número de grupos por valor de raio de vigilância possui informação útil à seleção dos raios candidatos, indicando qual faixa, quantos e quais valores devem ser analisados.

Para a produção desta curva, uma possibilidade é, de forma similar ao critério proposto em [1], definir os extremos da faixa de valores de raio pelos valores mínimos

e máximos da distância dos eventos. Uma forma mais apropriada, no entanto, para esta definição é basear-se na distância mais freqüente (moda) dos eventos [164], índice que é menos susceptível à eventos espúrios, comuns no ambiente de sonar, e melhor retrata a estrutura existente nos dados. Esta definição pode ser realizada através de frações da moda, que podem estar associadas a diferentes granularidades.

O critério proposto é resumido na Figura 5.4. É possível observar que, inicialmente, realiza-se uma seleção das frações da moda a serem consideradas, com as quais são calculados os valores de raio correspondentes. Para cada valor de raio, é produzido um agrupamento. Com base nos agrupamentos produzidos, realiza-se uma curva do número de grupos por valor de raio. Através desta curva, cabe definir quais granularidades serão avaliadas, processo realizado através de heurísticas, as quais podem considerar, se disponível, informação a priori sobre o problema, ou, na sua ausência, basear-se numa relação arbitrária entre o número de grupos criados e a quantidade de eventos disponíveis. Para o conjunto de sonar, um agrupamento foi produzido por classe (vide seção 5.1.5), e a definição das frações foi realizada pela inspeção desta curva, considerando a última estratégia, para a qual a menor fração da moda considerada buscou produzir um número de grupos de $\frac{1}{10}$ a $\frac{1}{15}$ do número de eventos de cada classe, resultando numa faixa de valores de raio corresponde à 0,5 a 5,0 vezes o valor da moda para as diferentes classes e modalidades de compactação avaliadas. Caso a curva produzida indique que a faixa não é apropriada aos objetivos do problema, ou seja, que as granularidades de interesse não foram contempladas, retorna-se ao estágio inicial; em caso contrário, realiza-se a seleção dos agrupamentos candidatos à análise. Para cada agrupamento expressivo, produz-se um par de conjuntos de projeto e avaliação e realiza-se o treinamento e avaliação de um classificador. Por fim, identifica-se o classificador de maior eficiência de generalização, o qual define o raio de vigilância mais adequado.

5.1.3 Agrupamento hierárquico

O agrupamento hierárquico adota uma filosofia operacional significativamente diferente do agrupamento seqüencial [1], produzindo uma família de agrupamentos ao invés de um único. Duas estratégias podem ser utilizadas para sua formação: a divisiva e a aglomerativa, a última escolhida para este trabalho por seu menor custo

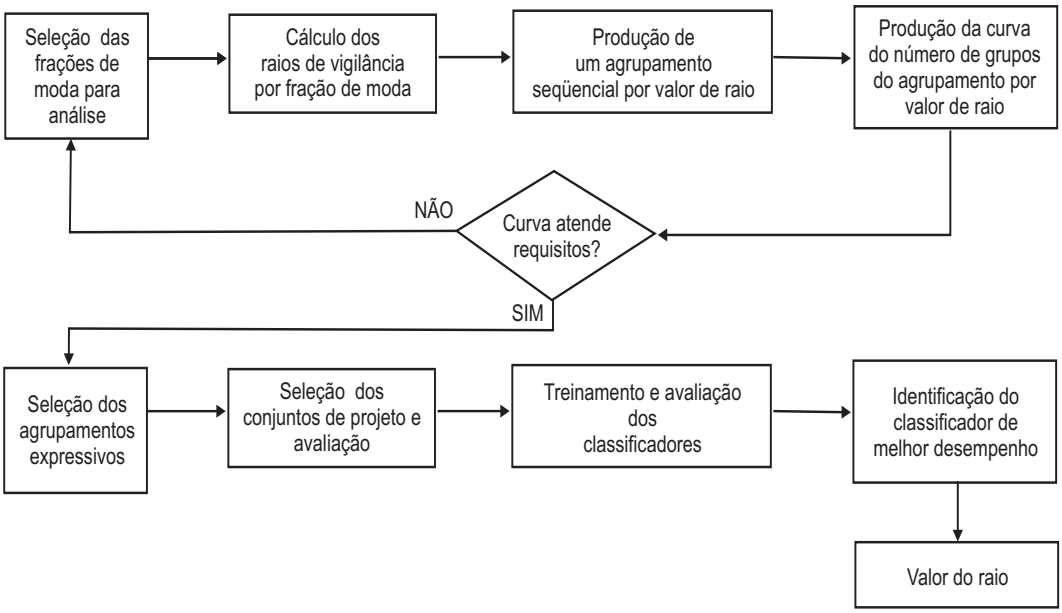


Figura 5.4: Fluxograma do algoritmo proposto para a seleção do valor do raio de vigilância do agrupamento seqüencial (Veja o texto).

computacional [163].

O processo de produção de um agrupamento hierárquico aglomerativo envolve vários estágios. No primeiro estágio, cada evento define um grupo. No estágio seguinte, dois grupos do estágio anterior são unidos para formar um mesmo grupo. Este processo é repetido até que todos os eventos pertençam a um único grupo, o que envolve $N-1$ estágios para um total de N eventos. Cada estágio define uma possível divisão dos dados. Uma desvantagem deste processo é que, se um grupo deficiente for formado no k-ésimo estágio, ele será mantido em todas as etapas posteriores [170].

Para cada estágio, é necessário identificar quais grupos serão unidos, o que envolve a avaliação de uma função de dissimilaridade para todas as fusões possíveis. Entre os vários candidatos, seleciona-se o par de grupos que apresentar um menor valor de dissimilaridade. Os principais critérios da literatura para a medida da dissimilaridade entre dois grupos R e Q $(s(R,Q))$ são resumidos na Tabela 5.1. Para a avaliação de similaridades entre os eventos, utilizam-se medidas de distância, conforme discutido na Seção 5.1.1.

É interessante observar que a função de dissimilaridade escolhida para a construção do agrupamento possui relação direta com a forma do agrupamento produzido. Para o critério de ligação simples, há uma tendência da formação de agupa-

Tabela 5.1: Critérios para a medida de dissimilaridade

Critério	**Medida**	**Referências**
Ligação simples	$s(R,Q) = \text{mín}\{s(i,j)\}$	Florek (1951) [171], Sneath (1957) [172]
Ligação completa	$s(R,Q) = \text{máx}\{s(i,j)\}$	McQuitty (1960) [173], Sokal (1963) [174]
Ligação média	$s(R,Q) = \frac{1}{N_R N_Q} \sum s(i,j)$	Michener (1958) [175]
Ward	$s^2(R,Q) = \frac{2N_R N_Q}{N_R+N_Q} \|\|\overline{x}(R) - \overline{x}(Q)\|\|^2$	Ward (1963) [176]

- N_X é o número de eventos do grupo X.

- $\overline{x}(X)$ é o centróide do grupo X.

- $s(i,j)$ é a dissimilaridade entre o i-ésimo evento de R e o j-ésimo evento de Q.

mentos alongados que sofrem o efeito referido como encadeamento (*channing effect* [1, 163])[2]. Para a ligação completa, a tendência é a formação de agrupamentos compactos, onde eventos similares tendem a permanecer em grupos distintos por vários estágios, efeito que é conhecido como dissecção [163]. As técnicas de ligação média e Ward apresentam uma menor sensibilidade a dados espúrios. Para o último critério, na formação de um novo grupo, são selecionados pares que minimizam sua variância, resultando em agrupamentos de mínima variância [1, 163].

O agrupamento hierárquico pode ser representado através de um gráfico referido como dendrograma [163]. Um dendrograma comum é o de dissimilaridade, que representa, a cada estágio, quais grupos foram unidos e o valor de dissimilaridade correspondente. Na Figura 5.5 é apresentado um dendrograma para 5 eventos arbitrários. Para este dendrograma, no primeiro estágio, forma-se um grupo envolvendo os eventos $\mathbf{x}_1$ e $\mathbf{x}_2$, para um valor de dissimilaridade de 2; no segundo estágio, tem-se um novo grupo, com $\mathbf{x}_4$ e $\mathbf{x}_5$, e um valor de dissimilaridade de 4; no terceiro, o evento $\mathbf{x}_3$ é anexado ao agrupamento formado no estágio dois, para uma dissimilaridade de valor 7; por fim, os agrupamentos do estágio um e três são unificados segundo uma dissimilaridade igual a 10. Um dos atrativos do dendrograma é prover um resumo

[2]Ainda que tenhamos dois grupos bem separados, caso haja apenas um par de eventos próximos entre os grupos, há uma tendência que ambos grupos sejam unidos.

desta modalidade de agrupamento, o qual permite avaliar, segundo a aplicação, se a união de um par de grupos num dado estágio é forçada ou natural [1].

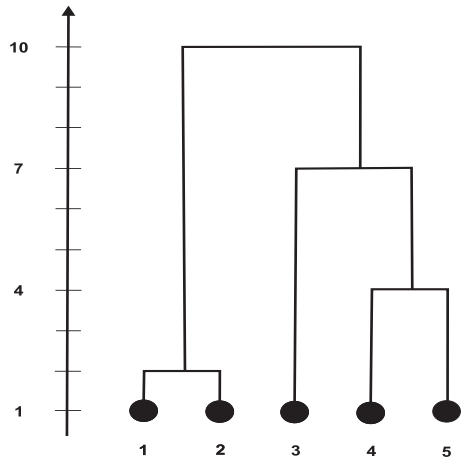


Figura 5.5: Exemplo de um dendrograma de dissimilaridade para dados arbitrários

Como o agrupamento hierárquico provê uma família de agrupamentos, é necessário escolher aquele que melhor retrata a estrutura existente nos dados. Para a seleção deste agrupamento, realiza-se o corte do dendrograma, especificando-se um nível máximo de dissimilaridade dos grupos, o que define o número de grupos existentes [177]. Para o exemplo da Figura 5.5, para um nível de corte de valor 6, os grupos identificados seriam: $\{1, 2\}$, $\{3\}$ e $\{4, 5\}$. Para um nível de corte de 8, os grupos passariam a ser: $\{1, 2\}$ e $\{3, 4, 5\}$.

5.1.3.1 Escolha do nível de corte

Conforme descrição anterior, para o agrupamento hierárquico é necessário realizar o corte do dendrograma. Comumente, este corte é realizado pelo especialista que, com base na inspeção visual do dendrograma e no conhecimento a priori dos dados, define o nível de corte mais apropriado [156]. Em geral, nesta inspeção são avaliados os valores de dissimilaridade envolvidos na formação dos grupos e identificadas fusões onde haja uma variação significativa deste valor, visto que a ocorrência de "saltos"de dissimilaridade sinaliza a integração de grupos bem determinados e distintos, que, preferencialmente, deveriam ser mantidos em grupos separados.

Para o problema em estudo, a produção de um agrupamento hierárquico por classe resultaria em dendrogramas com, no mínimo, 2142 níveis (classe G); e, no máximo, 7074 níveis, visto que um dendrograma para N eventos possui $(N - 1)$

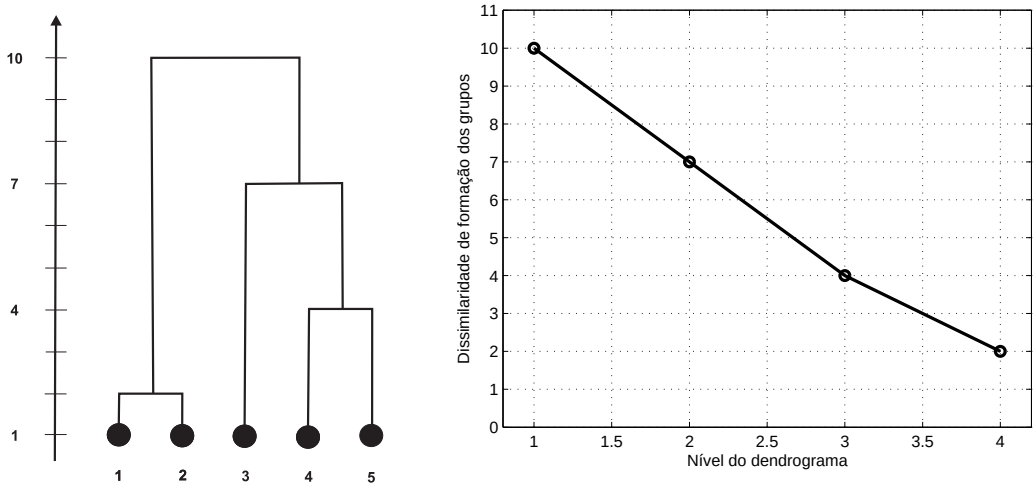
níveis. Não há, também, conhecimento a priori sobre os dados que possa orientar o processo de corte. Tendo em vista o elevado número de níveis existentes, a definição do nível de corte por inspeção visual é um processo complexo, portanto, contra-indicado.

Uma alternativa é realizar o corte do dendrograma através de um índice estatístico que identifique o número de grupos existentes. Esta solução, no entanto, de forma análoga ao agrupamento seqüencial (vide seção 5.1.2.1), pode resultar numa seleção dos conjuntos de projeto e avaliação inapropriada, pois pode se basear em agrupamentos não-realísticos. Entre índices aplicáveis ao agrupamento hierárquico, Halkidi et al.(2001) [178] propõe realizar o corte do dendrograma pela inspeção de uma curva produzida com base numa estatística que mede a dispersão dos grupos do agrupamento. O número de grupos é definido através da identificação de variações significativas do valor desta curva, processo que é bastante subjetivo.

Uma curva que pode fornecer um resumo útil do dendrograma é a da dissimilaridade de formação dos grupos versus nível do dendrograma. Na Figura 5.6 é ilustrada sua formação para um dendrograma arbitrário. Como o grupo formado no primeiro nível possui dissimilaridade igual a 10, o primeiro ponto é definido pelo par de coordenadas (1,10). De forma similar, os demais pontos seriam: $(2, 7)$, $(3, 4)$ e $(4, 2)$. Um de seus principais atrativos é fornecer, de forma compacta, um resumo das dissimilaridades envolvidas na formação dos diferentes grupos, o que possibilita identificar os "saltos"de dissimilaridade considerados no critério heurístico anteriormente mencionado, podendo, assim, orientar o processo de corte do dendrograma.

A dissimilaridade na formação de novos grupos está relacionada à granularidade dos agrupamentos propostos pelo algoritmo hierárquico. Tanto maior o nível de dissimilaridade de um novo grupo, menor é a granularidade considerada na constituição do agrupamento deste estágio, o que resulta num menor número de grupos. Deste modo, há uma relação inversa entre a granularidade e a dissimilaridade. Assim, a curva de dissimilaridade pode ser utilizada para a identificação de níveis de corte que produzem agrupamentos de diferentes granularidades, de forma similar à curva do número de grupos criados versus valor raio de vigilância, no agrupamento seqüencial (vide seção 5.1.2.1).

Visto que a curva de dissimilaridade pode orientar a seleção de um conjunto de



(a) (b)

Figura 5.6: Ilustração da curva de dissimilaridade de formação dos grupos versus nível do dendrograma: (a) dendrograma base e (b) curva produzida.

níveis de corte associados a diferentes granularidades, um critério possível é produzir diferentes cortes e utilizar o classificador como figura de mérito para a definição do corte mais apropriado, de forma análoga ao realizado com o agrupamento seqüencial.

Segundo qual critério, no entanto, serão definidos os níveis de corte? Como, em virtude do expressivo número de eventos e classes existentes, um agrupamento hierárquico será produzido para cada classe, é desejável que um mesmo critério seja utilizado para o corte das diferentes classes. No caso do agrupamento seqüencial, a moda das distâncias dos eventos unificou a seleção dos agrupamentos para as diferentes classes. Para o agrupamento hierárquico, faz-se necessário, no entanto, prover um critério que uniformize o corte dos dendrogramas das diferentes classes, levando em consideração a estrutura existente nos dados.

A observação das curvas de dissimilaridade relacionadas às diferentes classes, a serem apresentadas na seção 5.1.5.3, mostrou uma característica, tipicamente exponencial decrescente, o que motivou a proposição de um critério que explorasse a idéia de constantes de decaimento. Por este critério, as granularidades são associadas, abstratamente, a dado número de constantes de decaimento. Cabe observar que, para uma função exponencial do tipo: $y = ae^{-x}$, NC constantes de decaimento correspondem ao valor de x para o qual o valor inicial de y, dado por $y_0 = a$, reduz-se para $y_{NC} = ae^{-NC}$. Assim, uma granularidade mais grosseira, que resultaria na

seleção de agrupamentos com um menor número de grupos, estaria associada a um menor número de constantes de decaimento. Para uma granularidade mais fina, ocorreria o contrário.

Para identificar o corte associado a cada constante de decaimento definida no critério anterior, realiza-se uma modelagem da curva de dissimilaridade. Várias propostas de funções para esta modelagem foram avaliadas, sendo que a seguinte função apresentou bons resultados:

$$f(n) = e^{p_1+p_2 n+p_3 log(n)}, \tag{5.5}$$

onde n é o número de grupos do agrupamento, e p_1, p_2 e p_3 são constantes a serem determinadas por um *software* de ajuste de curvas (*fitting*). O valor de n associado a dado número de constantes de decaimento (NC) pode ser determinado por:

$$\frac{f(n)}{f(1)} = e^{-NC}, \tag{5.6}$$

ou seja, corresponde ao número de grupos (n) para o qual o valor inicial de dissimilaridade $f(1)$ reduz-se para $f(1)e^{-NC}$. Esta igualdade, se manipulada algebricamente, resulta na seguinte equação não-linear em n:

$$p_2 n + p_3 log(n) + NC - p_2 = 0, \tag{5.7}$$

cuja solução, para cada NC, pode ser determinada de forma numérica. Caso o valor de n retornado pela Equação 5.7 não seja inteiro, utiliza-se o inteiro mais próximo.

Com o corte definido pelo número de constantes de decaimento, cabe estabelecer um conjunto de constantes para serem avaliadas através do classificador. Para esta definição, o critério proposto utiliza uma curva do nível de corte ou do número de grupos do agrupamento selecionado como função do número de constantes de decaimento. Para esboçar esta curva, arbitra-se uma faixa e, através da Equação 5.7 que nos permite obter os parâmetros p_1, p_2 e p_3 do *fitting*, são obtidos os níveis de corte correspondentes. Esta curva permite identificar grupos de constantes de decaimento que resultam na seleção de agrupamentos com diferentes granularidades, de forma similar à curva do número de grupos versus fração da moda, considerada para o agrupamento seqüencial, conforme seção 5.1.2.1.

Para fins ilustrativos, na Figura 5.7(a), é exibida a curva de dissimilaridade de um agrupamento seqüencial hipotético, juntamente com sua curva ajustada, cujo

o modelo matemático produzido foi $y = e^{-0,5NC}$. Pela solução da Equação 5.7, $n = 2NC$, que define a curva do nível de corte por frações de constante de decaimento exibida na Figura 5.7(b).

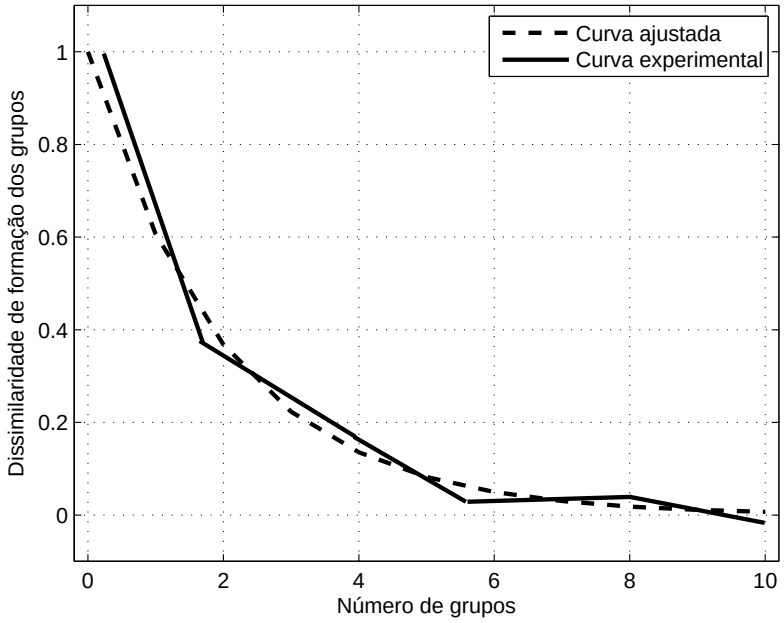
Na Figura 5.8 é apresentado um diagrama resumo do critério proposto. Para cada agrupamento considerado, produz-se uma curva de dissimilaridade e realiza-se o ajuste do modelo. Com base no modelo, é gerada a curva do número de grupos por constante de decaimento, com a qual são selecionadas as constantes de decaimento de interesse. De posse destas constantes, realiza-se o corte do dendrograma, sendo produzidos pares de conjuntos de projeto e avaliação. Para cada par, um classificador de mesma topologia é treinado e avaliado. Por fim, o nível de corte é definido pelo classificador de melhor desempenho.

5.1.4 Índices para a seleção de parâmetros

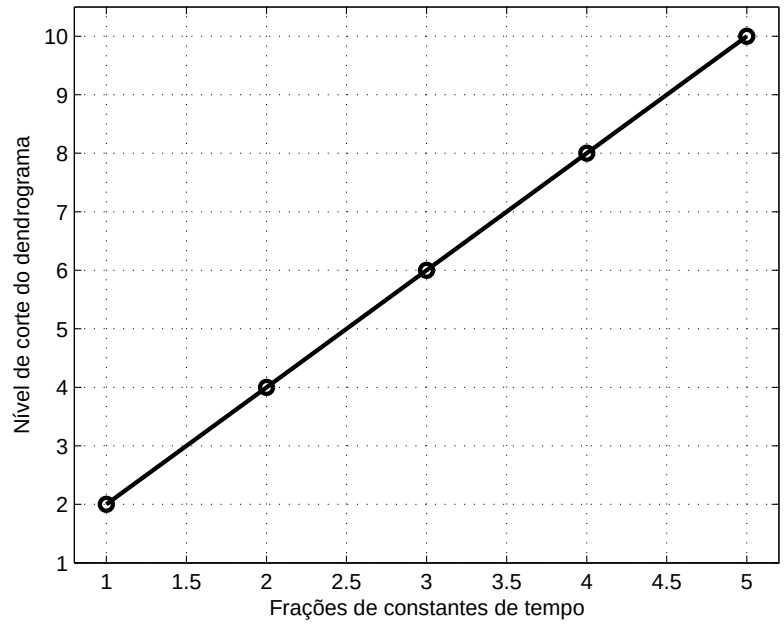
Uma alternativa para a seleção do raio de vigilância no agrupamento seqüencial, ou do nível de corte, no hierárquico, consiste em identificar, nestes agrupamentos, o provável número de grupos existentes.

Para esta identificação, um critério possível é produzir vários agrupamentos, cada qual utilizando um valor de parâmetro particular, e selecionar, através de um índice estatístico específico, qual dos agrupamentos melhor reflete a estrutura existente nos dados. Este critério de escolha é referido, na literatura, como relativo [1], visto que a definição do parâmetro é baseada na comparação de agrupamentos previamente produzidos.

Índices estatísticos para a identificação do número de grupos existentes nos dados usualmente avaliam as propriedades espaciais dos grupos identificados pelos agrupamentos. Um critério freqüentemente explorado consiste em privilegiar agrupamentos com grupos compactos e bem separados entre si. Há, no entanto, diferentes interpretações de como aferir a separação e a compactação dos grupos, resultando na proposição de diferentes índices pela literatura. Entre índices expressivos, tem-se: *Silhouette* [179], *Dunn* [159], *Davies e Bouldin* [160] e algumas variações, que serão discutidos a seguir.



(a)



(b)

Figura 5.7: Ilustração da curva de dissimilaridade de formação dos grupos (a) e da curva de nível de corte por número de constantes de decaimento (b) para o exemplo discutido no texto.

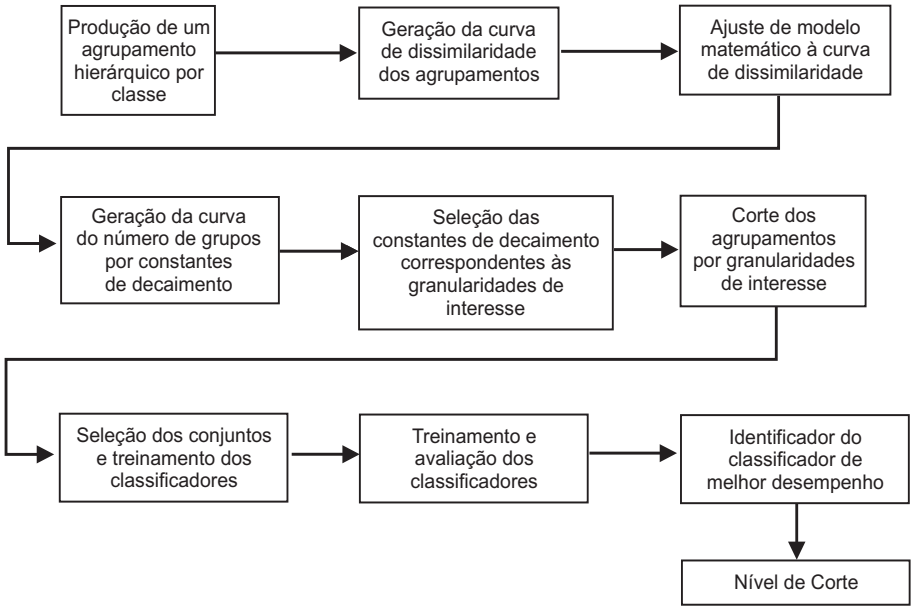


Figura 5.8: Diagrama em blocos do processo de corte do agrupamento hierárquico proposto (Veja o texto).

5.1.4.1 Índice de *Silhouette*

Introduzido por Rousseauw (1987) [179], o índice é baseado na comparação da dissimilaridade de cada evento em relação ao grupo que o contém (dissimilaridade intragrupo) com a dissimilaridade para os demais grupos existentes (dissimilaridade intergrupo).

Seja um agrupamento formado por m-grupos $\{C_1, \ldots, C_m\}$, cujos elementos são determinados pelo conjunto $\{S_1, \ldots, S_m\}$, onde S_j, para $1 \leq j \leq m$, define os índices dos eventos pertencentes ao j-ésimo grupo. Considere um evento i pertencente a um grupo j, logo $i \in S_j$. A dissimilaridade intragrupo associada a este evento é calculada como:

$$a(i) = \sum_{k \in S_j, k \neq i} d(i,k), \tag{5.8}$$

onde $d(i,k)$ representa a distância euclidiana entre dois eventos i e k definidos por S_j, logo pertencentes a C_j.

Analogamente, a dissimilaridade do evento i em relação a um grupo p, sele-

cionado dentre os $(m-1)$ demais grupos $(1 \leq p \neq j \leq m)$, é dada por:

$$c_p(i) = \sum_{k \in S_p} d(i,k) \tag{5.9}$$

Através de $c_p(i)$, a dissimilaridade intergrupo associada ao evento i é determinada como:

$$b(i) = \min\{c_p(i)\}, \tag{5.10}$$

ou seja, é definida pelo grupo de menor dissimilaridade, ou ainda, o grupo mais próximo do evento i. Para a determinação do coeficiente de Silhouette, é necessária ainda, para cada evento, a produção da seguinte regra:

$$s(i) = 1 - \frac{a(i)}{b(i)}, \quad a(i) < b(i) \tag{5.11}$$

$$s(i) = 0, \quad a(i) = b(i) \tag{5.12}$$

$$s(i) = \frac{b(i)}{a(i)} - 1, \quad a(i) > b(i), \tag{5.13}$$

para a qual: $-1 \leq s(1) \leq 1$. Esta estatística pode ainda ser reescrita na forma [163]:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}} \tag{5.14}$$

A estatística $s(i)$ possui uma interpretação interessante: caso seu valor seja próximo de $+1$, a dissimilaridade intergrupos é maior que a dissimilaridade intragrupos, e não há dúvida que o evento foi atribuído ao grupo correto. Para $s(i)$ próximo de -1, ocorre o contrário, e o evento está mais próximo de um grupo vizinho que do grupo a que foi atribuído. Para $s(i)$ próximo de zero, tem-se $a(i) \approx b(i)$, ou seja, não fica claro a que grupo o evento pertence. Conclui-se que o valor de $s(i)$ fornece uma medida da qualidade do agrupamento em relação ao i-ésimo evento. Para aferir a qualidade de todo um agrupamento, utiliza-se a largura média de silhouette para todo o conjunto de dados, a qual é determinada por:

$$\overline{s} = \frac{1}{N} \sum_{l=1}^{m} \sum_{k \in S_l} s(k), \tag{5.15}$$

onde N é o número total de eventos, sendo selecionado, dentre os agrupamentos em comparação, aquele que apresentar o maior valor de $\overline{s}$.

5.1.4.2 Índice de *Dunn*

Proposto por Dunn (1973) [159], este índice explora uma heurística similar ao índice de *Silhouette* [179], buscando agrupamentos compactos e bem separados, identificados através de medidas de disimilaridade intra e intergrupo. A dissimilaridade intragrupo é avaliada com base no diâmetro dos grupos. O diâmetro de um grupo X_i é definido como:

$$\Delta(X_i) = \max_{\mathbf{x},\mathbf{y}\in X_i}\{d(\mathbf{x},\mathbf{y})\}, \tag{5.16}$$

ou seja, corresponde a maior distância euclidiana verificada entre os eventos a ele pertencentes. Para a dissimilaridade intergrupo, é considerada a distância entre dois grupos X_i e X_j - $\delta(X_i, X_j)$ - dada por:

$$\delta(X_i, X_j) = \min_{\mathbf{x}\in X_i,\mathbf{y}\in X_j}\{d(\mathbf{x},\mathbf{y})\}, \tag{5.17}$$

que é definida pelo par de eventos mais próximo, considerando uma distância euclidiana, dos grupos X_i e X_j.

Para a produção do índice de *Dunn* será realizada uma medida baseada na razão das dissimilaridades inter e intragrupo. Esta razão, para dois grupos arbitrários X_i e X_j, é dada por:

$$s_{ij} = \frac{\delta(X_i, X_j)}{\max_{1\leq k\leq c}\{\Delta(X_k)\}}, \tag{5.18}$$

onde c corresponde ao número de grupos. O valor de s_{ij} é tanto maior quanto maior for a distância entre os grupos X_i e X_j e menor for o valor do diâmetro do "maior"grupo do agrupamento.

O índice de *Dunn* é determinado com base em s_{ij} pela fórmula:

$$I = \min_{1\leq i\leq c}\left\{\min_{1\leq j\leq c, j\neq i}\{s_{ij}\}\right\}, \tag{5.19}$$

ou seja, corresponde ao valor de d_{ij} associado ao par de grupos de menor valor de s_{ij}. Diferentemente do índice de *Silhouette*, que varia na faixa de -1 a 1, os valores apresentados pelo índice de *Dunn* são dependentes das características espaciais dos agrupamentos em análise.

Tanto maior o índice de *Dunn*, maior é a dissimilaridade intergrupo em relação a intragrupo, logo, dentre vários agrupamentos, seleciona-se aquele associado ao índice de maior valor.

5.1.4.3 Índice de *Davies e Bouldin*

Proposto por Davies e Bouldin (1977) [160], o índice considera uma dissimilaridade intragrupo ($s_{i,q}$) dada por:

$$s_{i,q} = \left(\frac{1}{N_{X_i}} \sum_{x \in X_i} ||\mathbf{x} - \mathbf{v}_i||^q \right)^{\frac{1}{q}}, \tag{5.20}$$

onde N_{X_i} é o número de eventos no conjunto X_i. A medida de dissimilaridade proposta pela Equação 5.20 considera o valor médio da distância de Minkowski [1] dos eventos ao centróide ($\mathbf{v}_i$) do grupo a que pertencem, onde q é um inteiro positivo arbitrário. Para a dissimilaridade intergrupos ($d_{ij,r}$) é utilizada a fórmula:

$$d_{ij,r} = \{ \sum_{s=1}^{p} |v_{si} - v_{sj}|^r \}^{\frac{1}{r}} = ||\mathbf{v}_i - \mathbf{v}_j||^r, \tag{5.21}$$

determinada, portanto, pela distância dos centróides dos grupos X_i e X_j, para r, também, inteiro positivo e arbitrário [3].

Produz-se ainda a medida $R_{ij,qr}$ dada por:

$$R_{ij,qr} = \frac{s_{i,q} + s_{j,q}}{d_{ij,r}}, \tag{5.22}$$

que relaciona a dispersão dos agrupamentos com a distância dos seus centróides. Posteriormente, para cada grupo X_i é produzida ainda a medida $R_{i,qr}$ dada por:

$$R_{i,qr} = \max_{j \neq i} R_{ij,qr}, \tag{5.23}$$

ou seja, é identificado outro grupo que forma com X_i um par tal que $R_{ij,qr}$ seja máximo. Se todos os grupos possuem uma dispersão similar, o par identificado por $R_{ij,qr}$ corresponde ao grupo mais próximo de X_i. O valor de $R_{i,qr}$ fornece, portanto, uma medida da contribuição de X_i para a dissimilaridade do agrupamento.

Para o cálculo do índice de *Davies e Bouldin* é necessário considerar a contribuição de cada grupo do agrupamento, logo o índice é definido por:

$$I_{qr} = \frac{1}{m} \sum_{i=1}^{m} R_{i,qr}, \tag{5.24}$$

[3]Note que as dissimilaridades intra e intergrupos, no cálculo de um mesmo índice, podem utilizar diferentes critérios de distância, ou seja, é possível que $q \neq r$. Freqüentemente, considera-se $q = r = 2$, valor para o qual a distância de Minkowski reduz-se à distância euclidiana.

onde m corresponde ao número de grupos existentes no agrupamento. Deste modo, para um conjunto de agrupamentos, quanto mais compactos e separados forem os grupos, menor será o índice. Assim, na seleção de um dentre um conjunto de agrupamentos disponíveis, considera-se aquele associado ao índice de menor valor.

5.1.4.4 Generalizações do Índice de *Dunn* e *Davies e Bouldin*

No cálculo do índice de *Dunn*, as medidas de dissimilaridade intra e intergrupo (Equações 5.16 e 5.17) são bastante sensíveis a presença de dados espúrios por se basearem em valores máximos e mínimos das distâncias. Assim, uma eventual inserção ou remoção de um único evento pode alterar, significativamente, ambas medidas. Para minimizar este problema, Bezdek e Pal (1998) propuseram generalizações deste índice, alterando o processo de cálculo das dissimilaridades envolvidas.

Para a medida de dissimilaridade intragrupo são propostos três critérios [162], apresentados na Tabela 5.2 [4], todos eles baseados na distância euclidiana dos eventos. Pode-se observar que o primeiro critério é o mesmo proposto por Dunn. Para o segundo, é utilizado o dobro da distância média entre os pares de eventos de X_i. Para o terceiro, tem-se o dobro da distância média dos eventos ao centróide ($\mathbf{v}_i$) do grupo. Como se baseiam em valores médios, o segundo e terceiro critérios são menos sensíveis a dados espúrios.

Tabela 5.2: Critérios para a medida da dissimilaridade intragrupo

Tipo	**Fórmula**
1	$\Delta_1(X_i) = \max_{\mathbf{x},\mathbf{y} \in X_i} \{d(\mathbf{x},\mathbf{y})\}$
2	$\Delta_2(X_i) = \frac{1}{N_{X_i}(N_{X_i}-1)} \sum_{\mathbf{x} \neq \mathbf{y}; \mathbf{x},\mathbf{y} \in X_i} d(\mathbf{x},\mathbf{y})$
3	$\Delta_3(X_i) = 2\left(\frac{\sum_{\mathbf{x} \in X_i} d(\mathbf{x},\mathbf{v}_i)}{N_{X_i}}\right)$

Quanto à dissimilaridade intergrupo, os seis critérios propostos [162] são apresentados na Tabela 5.3, os quais também são baseados em distância euclidiana. O primeiro critério corresponde ao proposto por *Dunn*. Para o segundo, o par de eventos mais distante define a dissimilaridade entre dois grupos. Para o terceiro,

[4] O número de eventos e o centróide do grupo X_i são dados por N_{X_i} e v_i, respectivamente.

esta dissimilaridade corresponde a distância média entre os eventos dos grupos. No quarto, é utilizada a distância dos centróides dos grupos. Para o quinto, é considerado um valor médio que envolve a distância dos eventos de X_i ao centróide de X_j, e vice-versa. O sexto envolve a métrica de *Hausdorff* [180] [5].

Tabela 5.3: Critérios para a medida de dissimilaridade intergrupo

Tipo	**Fórmula**
1	$\delta_1(X_i, X_j) = \min_{\mathbf{x}\in X_i, \mathbf{y}\in X_j}\{d(\mathbf{x}, \mathbf{y})\}$
2	$\delta_2(X_i, X_j) = \max_{\mathbf{x}\in X_i, \mathbf{y}\in X_j}\{d(\mathbf{x}, \mathbf{y})\}$
3	$\delta_3(X_i, X_j) = \frac{1}{N_{X_i} N_{X_j}} \sum_{\mathbf{x}\in X_i, \mathbf{y}\in X_j} d(\mathbf{x}, \mathbf{y})$
4	$\delta_4(X_i, X_j) = d(\mathbf{v}_i, \mathbf{v}_j)$
5	$\delta_5(X_i, X_j) = \frac{1}{N_{X_i}+N_{X_j}} \left(\sum_{\mathbf{x}\in X_i} d(\mathbf{x}, \mathbf{v}_i) + \sum_{\mathbf{y}\in X_j} d(\mathbf{x}, \mathbf{v}_j) \right)$
6	$\delta_6(X_i, X_j) = \max\{\delta(X_i, X_j), \delta(X_j, X_i)\}$ para: $\delta(X_i, X_j) = \text{máx}_{\mathbf{x}\in X_i}\{\text{mín}_{\mathbf{y}\in X_j}\{d(\mathbf{x}, \mathbf{y})\}\}$

Analogamente, o índice de *Davies e Bouldin* pode ser generalizado com base nas alternativas de distância propostas nas Tabelas 5.2 e 5.3. Bolshakova e Azuaje (2003) [181] utilizaram algumas generalizações deste índice para a análise de agrupamentos considerando dados provenientes do genoma humano. Nestas generalizações, o valor de $s_{i,q}$ é determinado por:

$$s_{i,q} = \Delta_k(X_i), \tag{5.25}$$

onde $\Delta_k(X_i)$ é selecionado entre os candidatos da Tabela 5.2, ou seja, $1 \leq k \leq 3$. O valor de $d_{ij,r}$ é calculado por:

$$d_{ij,r} = \delta_u(X_i, X_j), \tag{5.26}$$

para $\delta_u(X_i, X_j)$ selecionado entre as opções da Tabela 5.3, logo $1 \leq u \leq 6$.

[5] Nesta métrica, para cada ponto de X_i, seleciona-se o ponto mais próximo de X_j, formando-se um par. Entre os vários pares produzidos, é considerado àquele de maior distância, o qual define o valor de $\delta(X_i, X_j)$. Este procedimento é repetido invertendo-se a ordem dos conjuntos considerados. O valor da métrica é dado pelo maior valor obtido nos dois casos.

5.1.5 Resultados para a seleção estatística baseada em espectros

Em razão do expressivo número de eventos e classes disponíveis, e a fim de que o agrupamento melhor refletisse as especificidades de cada classe, foi produzido um agrupamento por classe. Esta produção, para a classe E, que possui 7075 eventos, exigiu, em virtude do custo computacional envolvido, a divisão de seus eventos em duas subclasses (E1 e E2), cada uma com metade das corridas disponíveis, o que resultou em subclasses com 3743 e 3332 eventos, respectivamente.

Conforme discutido na seção 5.1.1.1, fez-se necessário identificar qual técnica de compactação a ser aplicada aos espectros pré-processados seria a mais adequada. Nesta análise fez-se também necessário identificar o número de componentes a serem utilizadas, visto ser desejável reter o mínimo de componentes, mantendo ainda assim a informação relevante à classificação.

A seguir, será discutida a extração das componentes principais e de discriminação, o que definirá um conjunto de opções quanto a técnica de compactação e ao número de componentes a ser avaliado nesta seleção dos conjuntos. Após, serão apresentados resultados referentes à seleção dos conjuntos pelo agrupamento seqüencial, considerando-se os raios de vigilância escolhidos através de granularidades e de índices. Em seguida, será considerada a seleção baseada no agrupamento hierárquico. Por fim, os resultados produzidos pelas diferentes modalidades de seleção serão comparados, a fim de identificar a melhor partição da base de dados disponível ao projeto do classificador.

5.1.5.1 Compactação de espectros

Para a produção das componentes de representação (PCA) foi determinada a matriz de correlação dos dados, sendo extraídos seus autovalores e autovetores através de uma implementação do pacote *eispack* [115], amplamente utilizado pela comunidade científica. As componentes de discriminação foram extraídas por um classificador neural, cujo processo de treinamento é descrito no Apêndice B. Ambas modalidades de extração consideraram todo conjunto de dados disponível, visto que a partir das componentes extraídas é que serão gerados os agrupamentos utilizados na seleção dos conjuntos. O classificador extrator PCD foi treinado para um número

fixo de épocas, cujo valor foi determinado através de alguns ensaios.

Para a definição do número de componentes a serem utilizadas, considerando as componentes de representação, uma figura de mérito é a curva de acumulação de energia. Esta curva estabelece uma relação entre o número de componentes e a quantidade de energia do processo estocástico que é retida nos dados compactados, conforme discutido no Apêndice B. Nesta curva, esboça-se, no eixo vertical, a fração da energia total do processo que é retida por determinado número de componentes, o qual é identificada pelo eixo horizontal. Quanto mais acentuada a curva, maior é o poder de compactação, dado que um menor número de componentes reterá um valor de energia especificado. Esta curva, para o conjunto de dados (pré-processado) em estudo, é exibida na Figura 5.9, tendo como base os espectros pré-processados, com 557 componentes no total. É possível observar que a primeira componente retém quase 60% da energia total do processo. Este valor cresce para em torno de 68% e de 72% para 5 e 10 componentes, atingindo 90% para 140 componentes.

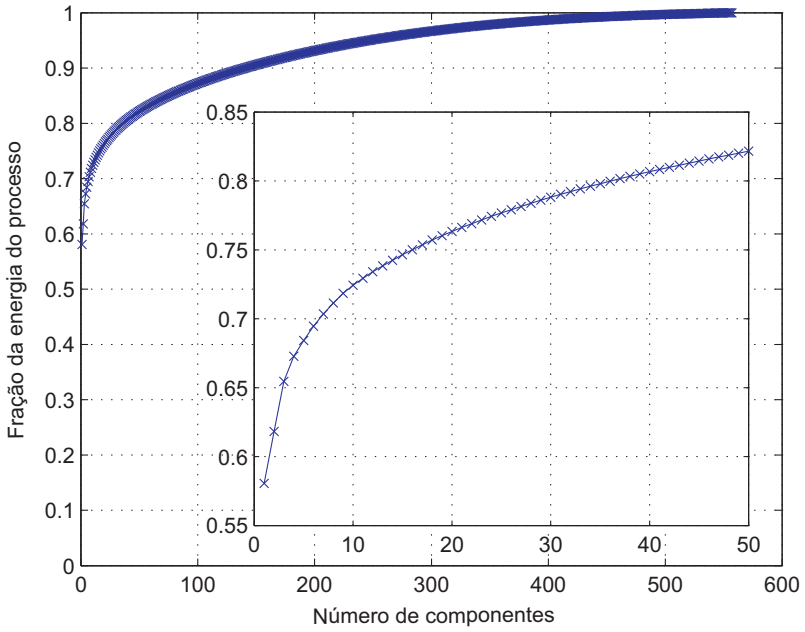


Figura 5.9: Curva de acumulação de energia para a análise PCA.

Em relação à extração de componentes de discriminação, com base no próprio treinamento do classificador extrator, pode-se produzir um gráfico das eficiências de

classificação obtidas para um número crescente de componentes. Esta curva fornece uma estimativa da eficiência da classificação neural dos dados compactados com base nestas direções, podendo ser utilizada como uma medida qualitativa da quantidade de informação discriminante que é retida num dado número de componentes. Na Figura 5.10, os valores médios da eficiência SP, juntamente com seus desvios associados, estimados pelo critério discutido na seção 4.2.1, são apresentados para um número crescente de componentes, considerando a faixa de 3 a 50 componentes, visto que para 1 e 2 componentes, o classificador não é capaz de classificar corretamente uma ou mais classes, o que resulta numa eficiência SP nula. Verifica-se que o crescimento é mais acentuado até atingirmos 15 componentes, reduzindo-se na faixa de 15 a 30 componentes. A partir de 20 componentes, tem-se um crescimento pouco significativo. Em termos de valores médios da eficiência SP e suas incertezas, para 3 componentes, tem-se um valor de $(64,4 \pm 1,4)\%$. Este valor sobe para $(82,6 \pm 0,4)\%$ e $(88,5 \pm 0,3)\%$ para 4 e 5 componentes, atingindo o valor de $(93,8 \pm 0,2)\%$ para 10 componentes. Para o último, tem-se uma eficiência mínima de $(88,5 \pm 0,4)\%$ (classe C) e máxima de $(97,7 \pm 0,3)\%$ (classe F) por classe.

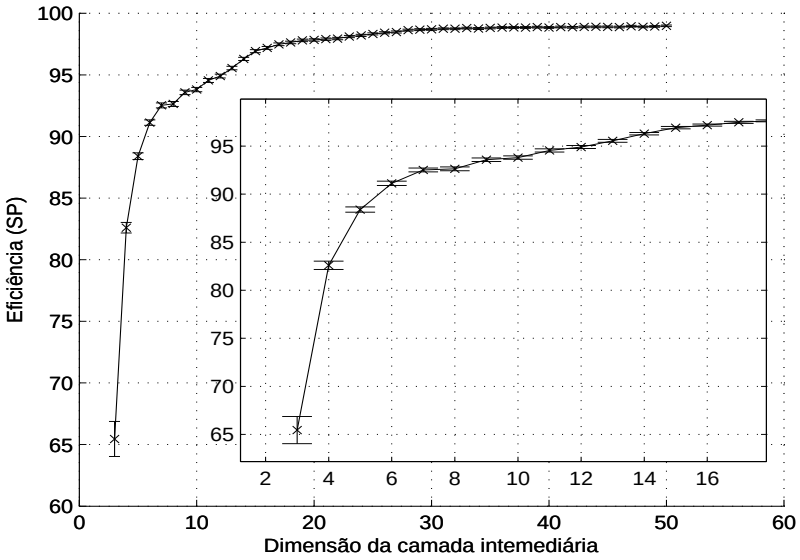


Figura 5.10: Curva de eficiências SP da rede extratora PCD para um número crescente de componentes.

Aspecto desejável em qualquer compactação é reter, num mínimo de componentes, o máximo de informação relevante ao problema, a qual, para o problema do sonar, deve ser discriminante. Deste modo, um procedimento natural é basear-se nas componentes de discriminação e, com base na curva de eficiência, definir o número de componentes. Para esta definição, identifica-se o número mínimo de componentes de discriminação que representa um bom compromisso entre complexidade e eficiência de classificação. Com base neste número, geram-se os espectros compactados utilizando ambas componentes de representação e discriminação, realizam-se os agrupamentos e produzem-se os classificadores. O classificador de melhor desempenho identifica a compactação mais apropriada, isto é, a que produziu os agrupamentos mais adequados para a seleção dos conjuntos de projeto e teste.

Através da Figura 5.10, foi identificado que 5 e 10 componentes representavam compromissos interessantes entre complexidade e eficiência. Deste modo, nesta análise inicial, 4 modalidades de compactação foram selecionadas para avaliação futura: duas considerando 5 e 10 componentes de representação (PCA-5C e PCA-10C), e duas outras para 5 e 10 componentes de discriminação (PCD-5C e PCD-10C). Acredita-se, portanto, que em 5 ou 10 componentes haja informação suficiente para diversos níveis de produção e avaliação dos agrupamentos.

A fim de verificar, qualitativamente, a quantidade de informação discriminatória retida por cada modalidade de compactação, foram treinados classificadores neurais (MLP) com base em dados compactados (PCA-10C, PCD-5C e PCD-10C), utilizando todos dados disponíveis, isto é, sem figuras de mérito para a generalização do treinamento. Na Figura 5.11 são apresentadas, para cada proposta, as eficiências SP e os desvios correspondentes, considerando diferentes números de neurônios na camada intermediária do classificador.

É possível observar que, até 2 neurônios, o melhor desempenho é da compactação PCD-5C. Para 3 ou mais neurônios, a compactação PCD-10C é a melhor. Em toda faixa, a compactação PCA-10C resulta em classificadores de menor eficiência que as obtidas por PCD. Há uma tendência de estabilização nas curvas para todas as compactações: em torno de 10 neurônios, com eficiência SP de $(86,7 \pm 0,3)\%$, para PCD-5C; próximo a 20 neurônios, segundo uma eficiência SP de $(94,8 \pm 0,2)\%$, para PCD-10C; e na vizinhança de 35 neurônios, com eficiência SP de $(85,5 \pm 0,4)\%$,

para PCA-10C. Esta estabilização indica que não há ganho expressivo de eficiência com a inserção de novos neurônios, podendo ser utilizada como indicativo de termos atingido uma complexidade adequada para o classificador. Assim, a compactação baseada em componentes de discriminação tendeu a apresentar modelos menos complexos, o que é um indicativo que a informação relevante à classificação é disponibilizada em maior volume e/ou de forma mais simples. Para as componentes de discriminação, o aumento do número de componentes (5 para 10) elevou a complexidade indicada como adequada para o classificador, o que acredita-se estar relacionado ao maior volume de informações discriminates disponíveis nos dados compactados.

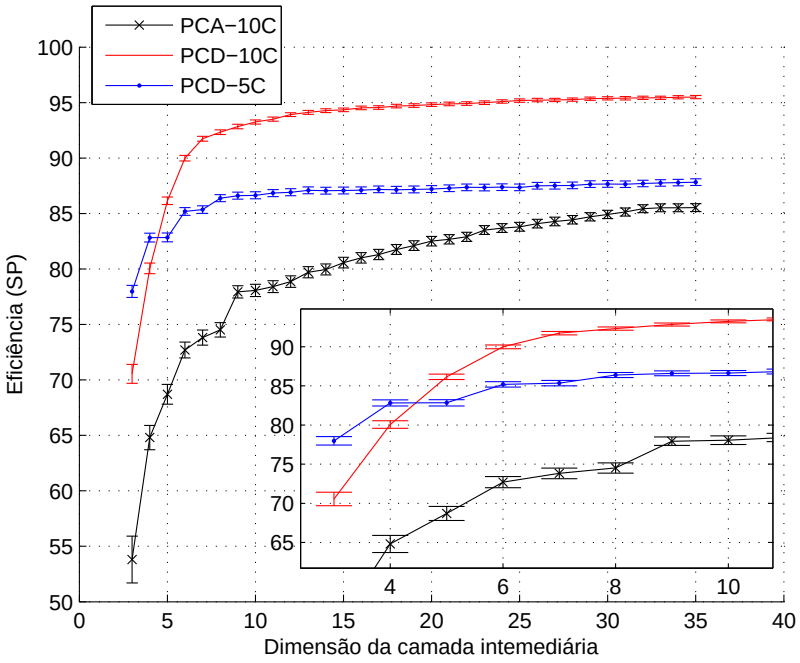


Figura 5.11: Curva de eficiências do classificador sobre espectros compactados pelas análises PCA-10C, PCD-10C e PCD-5C.

5.1.5.2 Agrupamento seqüencial

A seleção dos conjuntos de projeto e avaliação baseados em espectros considerou duas possibilidades quanto à seleção dos raios de vigilância: através do critério proposto na seção 5.1.2.1, e através dos índices descritos na seção 5.1.4. Para ambos

critérios, faz-se necessário definir um conjunto de valores candidatos do raio, do qual será identificado o mais adequado.

Para definição do conjunto de valores de raio, um procedimento útil é produzir a curva do número de grupos criados por valor de raio, conforme descrito na seção 5.1.2.1. Como, na seção anterior, foram introduzidos quatro possíveis modos de compactação dos espectros, isto se traduz numa curva por classe e modalidade de compactação, o que resulta num total de 32 curvas.

A geração destas curvas exige que uma faixa de valores seja arbitrada para o raio. Com base em alguns trabalhos [2, 164] e ensaios, foi verificado que uma faixa de valores de raio adequada ao problema de sonar compreende valores de 0,5 a 5 vezes o valor da moda. Nesta faixa, o maior número de grupos criados situou-se em torno de 300, para classes com 3432 (classe B) e 4797 eventos (classe C), considerando a compactação em 10 componentes de discriminação e raio de 0,5 vezes a moda. O menor número de grupos encontrado foi de 2, obtido para um raio de 5 vezes a moda, nas classes F (PCD-10C) e C (PCA-10C).

Na Figura 5.12 são exibidas as curvas do número de grupos criados, considerando 12 valores de moda ($0,50;\ 0,75;\ 1,00;\ 1,25;\ 1,50;\ 2,00;\ 2,50;\ 3,50;\ 4,00;\ 4,50;\ 5,00$) selecionados na faixa arbitrada, para componentes de representação (Figura 5.12(a)) e discriminação (Figura 5.12(b)). É possível perceber que o número de grupos é dependente da classe e da modalidade de compactação. Em relação à compactação por componentes de discriminação (PCD), as curvas correspondentes a 5 e 10 componentes são, muitas das vezes, similares. Para as classes C, D, E1 e G, a compactação em 10 componentes apresentou um número maior ou igual de grupos que em 5 componentes. Para a classe E2, ocorre o contrário. Quanto às classes A e B, inicialmente, há um maior número de grupos para 10 componentes, ocorrendo o contrário para valores de raio superiores a 1.0.

Para as componentes de representação (PCA), há uma maior distinção entre o número de grupos criados para 5 e 10 componentes, quando comparada com as componentes de discriminação, em especial, para as classes C, E1 e E2. Para classes A, B, C, D, E1, E2 e H, a compactação em 5 componentes apresenta um maior número de grupos. Nas classes F e G, a compactação com maior número de grupos é dependente do valor do raio.

Quando comparadas às curvas referentes à compactação em 10 componentes de representação e discriminação, há um maior número de grupos na última para a maioria das classes (A,B,C,E2,H). Fato contrário ocorre para as classes E1, F e G. Para a classe D, o número de grupos criados para as duas modalidades de compactação é equivalente.

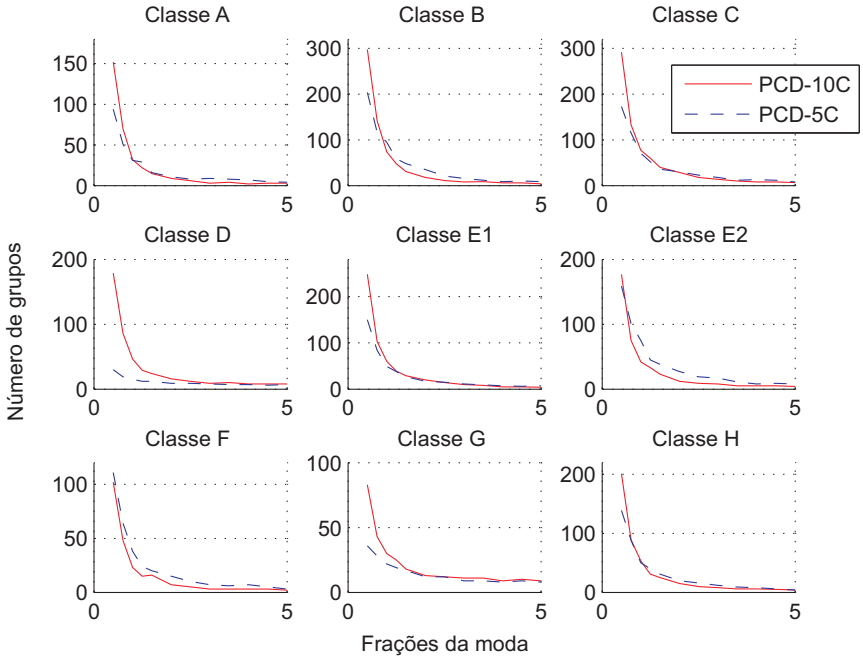
Portanto, as curvas sinalizam que a compactação em componentes de discriminação tende a produzir espaços de dados mais complexos que, possivelmente, podem reter um maior volume de informação dos dados, sendo assim mais interessantes para o processo de seleção dos conjuntos.

5.1.5.2.1 Seleção do raio de vigilância por granularidade

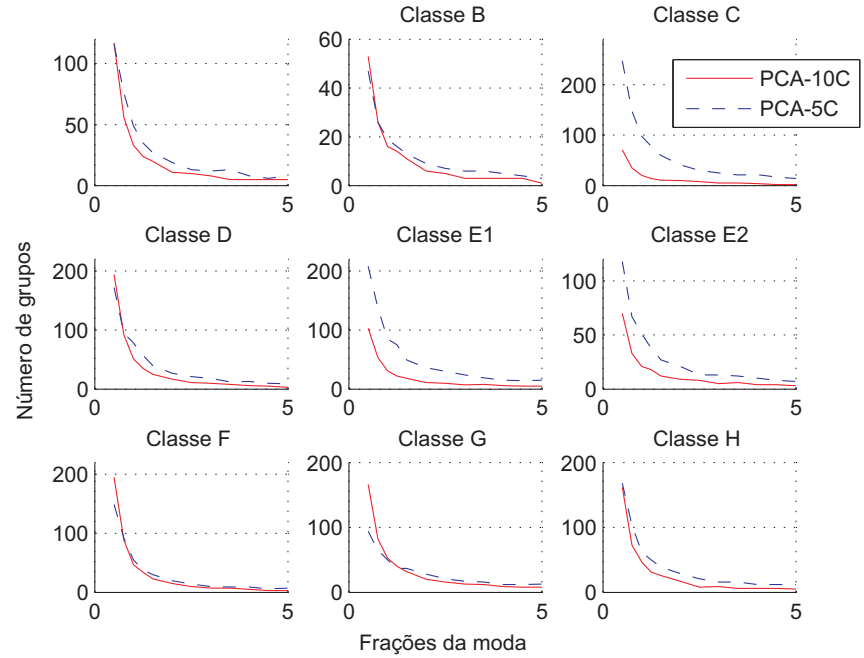
Pelo critério proposto, definida uma faixa de valores, cabe a seleção de um conjunto de raios expressivos, isto é, que resultem em agrupamentos de diferentes granularidades. Como a avaliação de cada raio deste conjunto demanda o treinamento de um classificador, o qual, para o problema do sonar, possui custo computacional alto, esta análise restringiu-se a três granularidades: uma mais fina, outra intermediária e, por fim, uma mais grosseira. A estas granularidades foram arbitrados os valores de raio de $\{0,5; 1,0; 5,0\}$ vezes a moda, aqui referidos como pequeno (P), intermediário (I) e grande (G), respectivamente, para as quais se verifica uma variação expressiva do número de grupos do agrupamento.

Visto que há um agrupamento para cada classe, a avaliação de todas as combinações possíveis destas três granularidades resultaria no treinamento de $3^8 = 6561$ classificadores, o que é proibitivo para o conjunto de sonar. Deste modo, optou-se por considerar uma mesma granularidade para todas as classes, o que resultou na avaliação de 12 grupos de classificadores, 3 para cada modalidade de compactação. Cada grupo considerou classificadores com um número de neurônios na camada intermediária escolhido na faixa de 10 a 40, a qual acredita-se permitir uma apropriada avaliação qualitativa das modalidades de seleção.

As eficiências SP dos classificadores associados às componentes de discriminação é apresentada na Figura 5.13. Através da Figura 5.13(a), o melhor desempenho para 10 componentes de discriminação é obtido para um valor de raio pequeno. Quando comparados os resultados referentes aos raios grande e pequeno, o último apresenta desempenho igual ou superior ao primeiro em quase toda faixa de neurô-



(a) Componentes PCD



(b) Componentes PCA

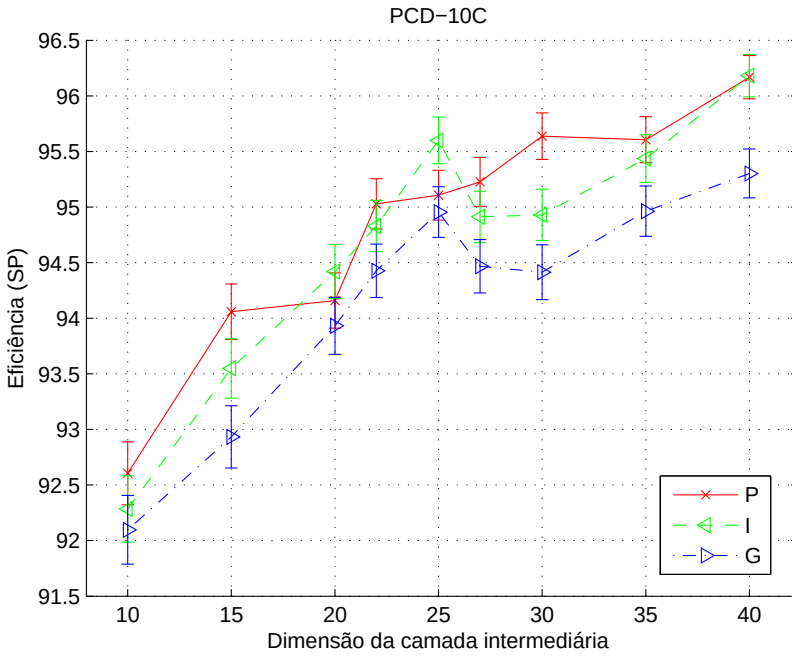
Figura 5.12: Número de grupos produzidos no agrupamento seqüencial para cada modalidade de compactação, raio inicial e classe de contato.

nios, apresentando diferenças de, aproximadamente, $0,2$ (22 neurônios) a $1,2$ (30 neurônios) pontos percentuais. Em relação ao raio intermediário, seus resultados são superiores ao raio grande, ainda que equivalentes ou inferiores ao raio pequeno para a maior parte dos ensaios, exceto para as redes com 20 e 25 neurônios. Na comparação entre raio pequeno e intermediário, as diferenças são de $\approx 0,2$ a $\approx 1,0$, excluídas as redes com 20 e 25 neurônios, para as quais há uma vantagem do raio intermediário de $\approx 0,2$ e $\approx 0,1$ pontos percentuais, respectivamente.

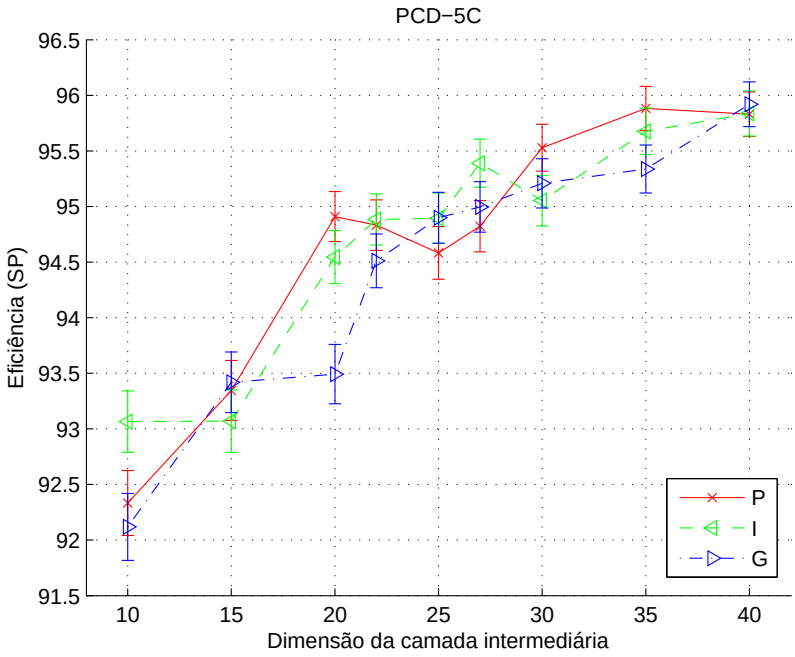
Para 5 componentes de discriminação, pela Figura 5.13(b), percebe-se que as diferenças são menos acentuadas, ainda que seja verificado um desempenho igual ou superior do raio pequeno em relação aos demais, exceto para as redes com 25 e 27 neurônios, onde o raio grande destaca-se em relação aos demais (diferenças de $\approx 0,4$ e $\approx 0,2$ pontos, respectivamente). A maior diferença de desempenho entre raio pequeno e grande ocorre para 20 neurônios e vale $\approx 1,5$ pontos percentuais. Para os raios pequeno e médio, a maior diferença é de $\approx 0,4$ para 35 neurônios.

Para 10 componentes de representação, conforme Figura 5.14(a), há um melhor desempenho para o raio intermediário. Na comparação entre raio intermediário e grande, o primeiro é melhor em toda faixa, apresentando diferenças de $\approx 0,1$ (20 neurônios) a $\approx 1,4$ (15 neurônios) pontos percentuais. Em relação ao raio pequeno, o intermediário possui melhor desempenho em quase toda faixa, apresentando diferenças, no entanto, menores, entre $\approx 0,05$ (27 neurônios) e $\approx 0,6$ (15 neurônios). Resultado similar é obtido para 5 componentes (Figura 5.13(b)), cujas diferenças entre os raios intermediário e grande estão entre $\approx 0,5$ (22 neurônios) e $\approx 1,5$ (10 neurônios). Entre os raios intermediário e pequeno, as diferenças situam-se na faixa de $\approx 0,25$ (22 neurônios) a $\approx 1,5$ (27 neurônios).

Na Figura 5.15 são comparadas as curvas associadas aos raios de melhor desempenho, segundo análise anterior. Pela Figura 5.15(a), para a qual são consideradas componentes de discriminação, há um melhor desempenho da compactação em 10 componentes, que apresenta uma vantagem de $\approx 0,1$ (30 neurônios) a $\approx 0,7$ (15 neurônios) pontos percentuais sobre a compactação em 5 componentes, exceto para 20 e 35 neurônios. Para as componentes de representação, em 5 dos 9 classificadores, há um desempenho superior da PCA-5C, com vantagens de $\approx 0,1$ (10 neurônios) a $0,7$ (15 neurônios) sobre a PCA-10C. A comparação entre as eficiências associadas

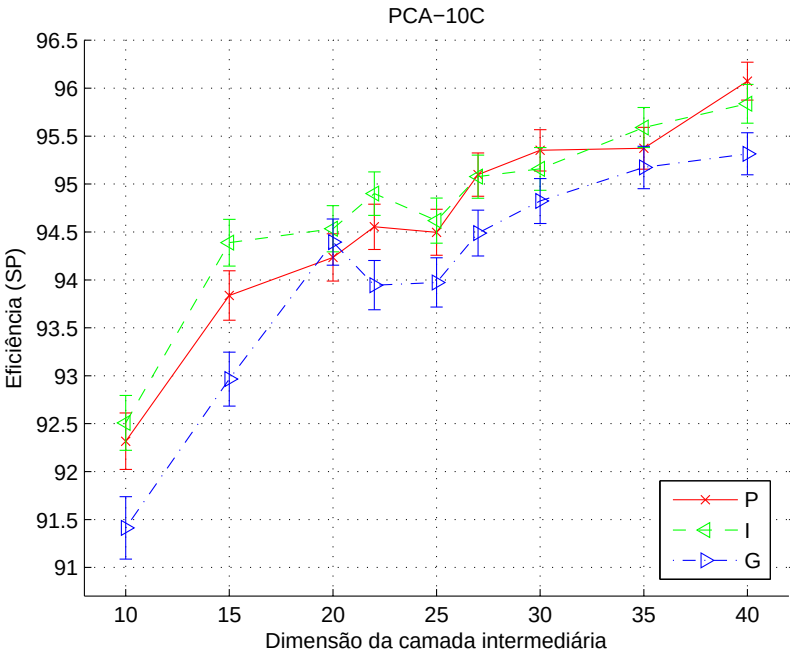


(a) 10 componentes de discriminação

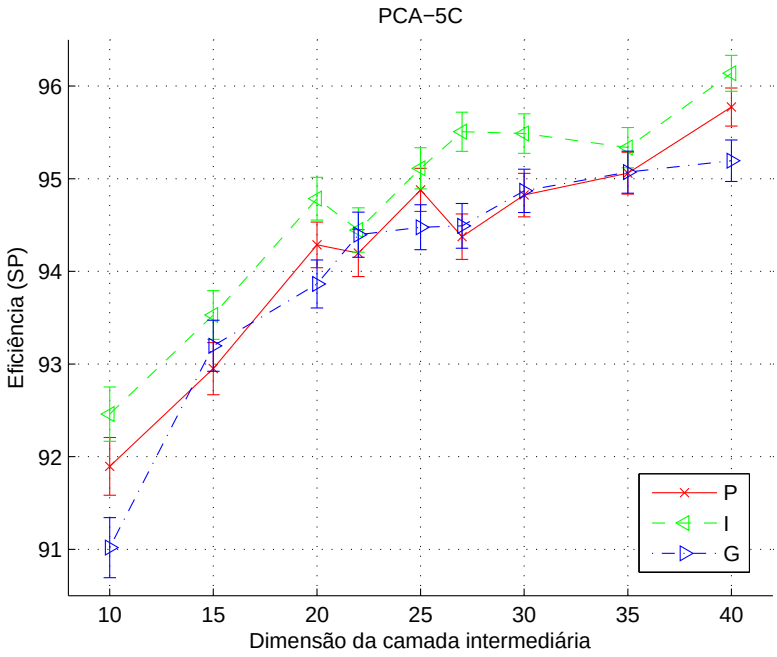


(b) 5 componentes de discriminação

Figura 5.13: Eficiências de generalização (SP) para o agrupamento seqüencial de espectros, considerando diferentes números de componentes de discriminação e valores de raio de vigilância.



(a) 10 componentes de representação



(b) 5 componentes de representação

Figura 5.14: Eficiências de generalização (SP) para o agrupamento seqüencial de espectros, considerando diferentes números de componentes de representação e valores de raio de vigilância.

às compactações de melhor desempenho para as componentes de discriminação e representação é realizada na Figura 5.16. Nesta comparação, há um desempenho equivalente ou superior da compactação em 10 componentes de discriminação, a qual apresenta uma vantagem até $\approx 0,5$ (15 neurônios) pontos percentuais, exceto para as redes com 20 e 27 neurônios.

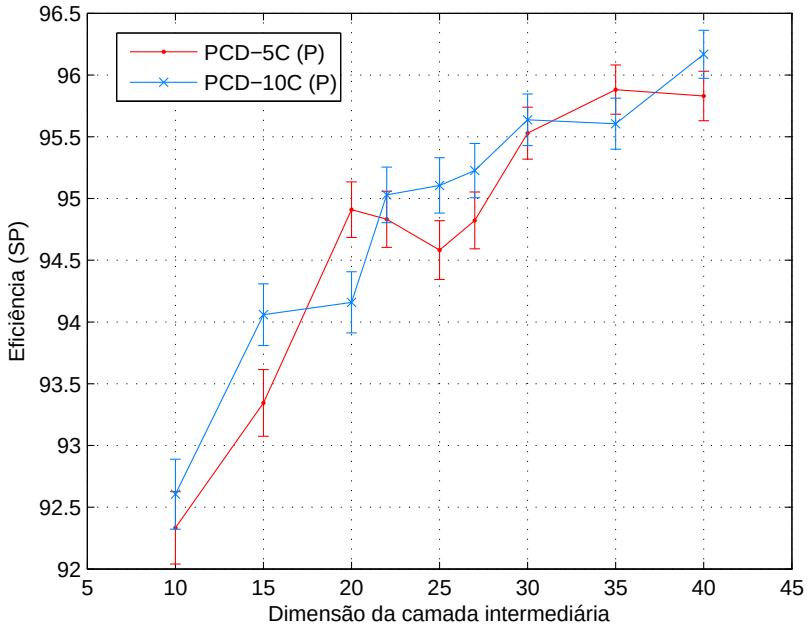
Dos resultados anteriores, para as componentes de discriminação, o melhor resultado do raio pequeno sinaliza uma granularidade mais fina para o tratamento dos dados compactados. Para as componentes de representação, tem-se uma granularidade média. Este resultado indica que os dados compactados com base em componentes de discriminação possuem uma estrutura mais complexa, resultado que é coerente com a discussão baseada no número de grupos criados, realizada anteriormente.

Pode-se perceber que a escolha do raio possui impacto no processo de seleção, ainda que as diferenças apresentadas mostraram-se inferiores à $1,5$ pontos percentuais, no pior caso. Esta diferença é, no entanto, expressiva, tendo em vista a complexidade do problema e as exigências quanto a confiabilidade e desempenho do sistema, o qual influencia decisões que envolvem risco de vida.

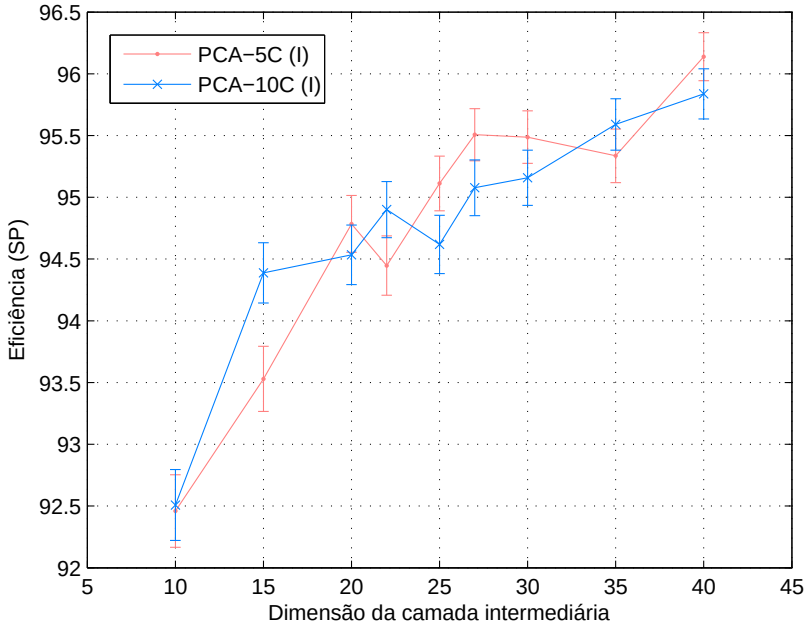
Nesta comparação, foram verificados casos onde o raio identificado como de melhor desempenho alterou-se de acordo com o número de neurônios da camada intermediária do classificador. Acredita-se que estas variações estejam relacionadas a flutuações inerentes ao processo de treinamento, que se mostraram da ordem das diferenças de desempenho verificadas para as diferentes escolhas dos raios e da modalidade de compactação. Tendo em vista estas flutuações, a escolha do raio de melhor desempenho foi sempre baseada no comportamento apresentado pelos classificadores para toda a faixa de neurônios considerada.

5.1.5.2.2 Seleção do raio de vigilância por índices

A avaliação da seleção dos valores de raio por índices considerou as 4 modalidades de compactação discutidas na seção 5.1.5.1 e os 12 valores de raio utilizados para o esboço da curva do número de grupos criados por valor de raio. Os índices selecionados foram: o *Silhouette* (SH), o *Davies-e-Bouldin* (DB) e o *Dunn* (DUNN), discutidos na seção 5.1.4. Para os dois últimos foram consideradas generalizações das distâncias intragrupo e intergrupo pelas propostas Δ_2 (Tabela 5.2) e δ_3 (Tabela



(a) componentes de discriminação



(b) componentes de representação

Figura 5.15: Eficiências de generalização (SP) para o agrupamento seqüencial de espectros, considerando componentes de representação e discriminação.

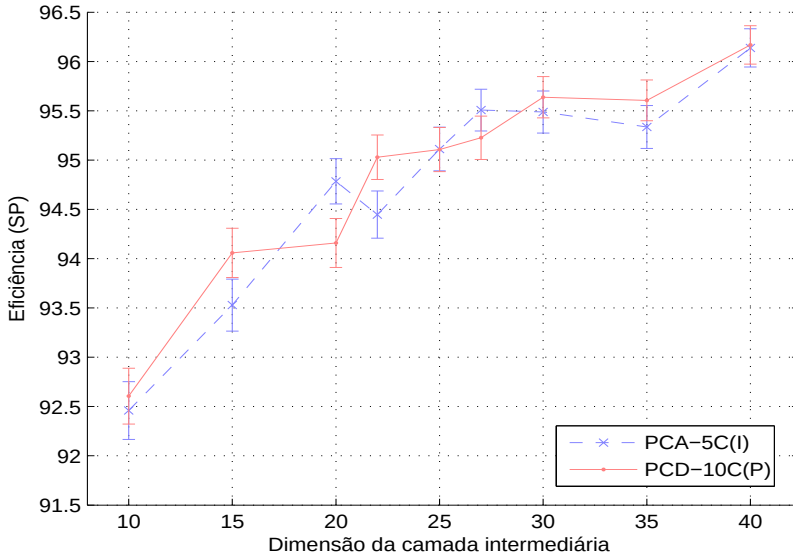
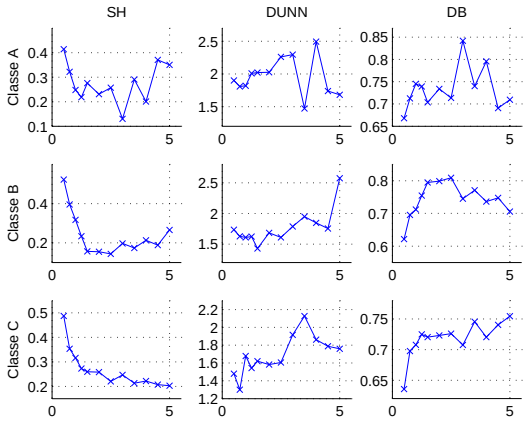


Figura 5.16: Eficiências de generalização (SP) para o agrupamento seqüencial de espectros, considerando 5 e 10 componentes de representação e discriminação, respectivamente.

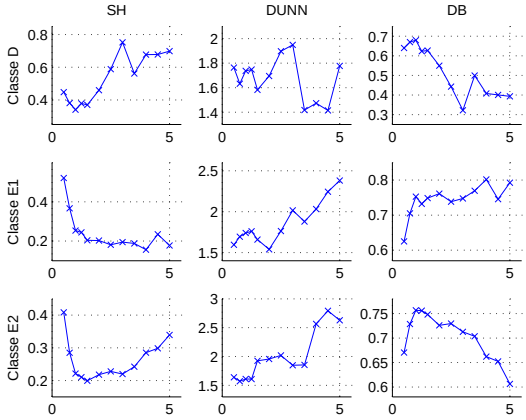
5.3), respectivamente, em razão de uma maior imunidade a dados espúrios.

Foram produzidos, para cada modalidade de compactação, classe e índice, gráficos relacionando os valores do índice (eixo vertical) com os multiplicadores da moda (eixo horizontal) . De posse dos gráficos, os pontos de máximo determinaram, pelos índices de *Silhouette* e *Dunn*, os melhores valores de raio. Para o índice de *Davies-e-Bouldin* foram considerados os pontos de mínimo. Por restrições de espaço, serão reproduzidas apenas as curvas relativas à compactação em 10 componentes de discriminação, que são apresentadas na Figura 5.17. Na Tabela 5.4 são resumidos os valores de raio selecionados para cada índice, classe e modalidade de compactação.

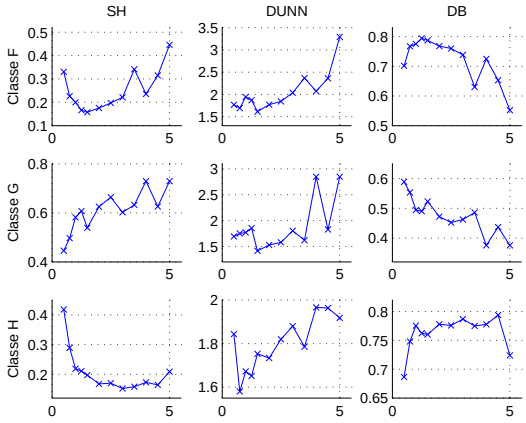
Analisando a Tabela 5.4, considerando uma compactação em 10 componentes de discriminação, é possível perceber uma total concordância entre os três índices para as classes D e F. Para a classe G, há uma concordância entre o SH e o DUNN (valor de raio 5,0), e o índice DB sinaliza em favor de um valor de raio 4,0. Para as classes A, B, C, E1 e H, os índices DB e SH apontam um valor de raio de pequeno (0,5); enquanto o índice DUNN, para um valor entre médio e grande (3,5 - 5,0). Para a classe E2, os índices DB e DUNN sinalizam um valor de raio grande (4,5-5,0); enquanto o SH, em favor de um raio pequeno (0,5). Em relação as componentes



(a)



(b)



(c)

Figura 5.17: Valores dos índices de *Davies e Bouldin* (DB), *Dunn* (DUNN) e *Silhouete* (SH) para os agrupamentos produzidos com base na compactação em 10 componentes de discriminação, para diferentes valores de raio inicial.

Tabela 5.4: Valores de raio inicial propostos pelos índices por proposta de compactação e classe

PCD-10C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
DB	0,5	0,5	0,5	3,0	0,5	5,0	5,0	4,0	0,5
SH	0,5	0,5	0,5	3,0	0,5	0,5	5,0	5,0	0,5
DUNN	4,0	5,0	3,5	3,0	5,0	4,5	5,0	5,0	4,0
PCA-10C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
DB	5,0	3,5	5,0	4,5	4,5	5,0	5,0	4,0	5,0
SH	4,5	3,5	5,0	4,5	4,5	5,0	5,0	4,0	5,0
DUNN	5,0	3,5	5,0	5,0	4,5	5,0	4,5	4,5	5,0
PCD-5C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
DB	5,0	0,5	3,5	4,5	0,5	4,0	5,0	4,0	5,0
SH	5,0	0,5	0,5	4,0	0,5	4,0	5,0	3,5	0,5
DUNN	4,5	5,0	4,5	4,5	5,0	5,0	0,5	3,5	2,0
PCA-5C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
DB	4,5	3,5	5,0	4,5	4,5	5,0	4,5	4,5	5,0
SH	4,5	3,5	5,0	4,5	4,5	5,0	4,5	4,0	5,0
DUNN	4,5	3,5	5,0	4,5	4,5	4,0	4,5	4,5	5,0

de representação, as compactações de 5 e 10 componentes apontam, para todas as classes, um valor de raio grande (4,5-5,0), exceto para a classe B, para a qual sinaliza-se um valor entre médio e grande (3,5). Uma maior divergência dos índices é verificada para 5 componentes de discriminação, onde valores de raio grandes (4,00-5,00) são verificado para as classes A, D e E2. Para a classe B e E1, os índices DB e SH apontam para um raio pequeno (0,5), enquanto o DUNN para um raio grande. Fato oposto ocorre para a classe F.

Para algumas modalidades de compactação, em especial, em componentes de discriminação, os três índices não se mostraram unânimes na escolha do valor do raio. Para estes casos, duas situações foram verificadas: uma primeira, onde os três índices sinalizaram valores de raio próximos, associados, portanto, a uma mesma granularidade (grosseira ou fina); e uma segunda, onde diferentes granularidades foram sinalizadas por cada índice. No último caso, para definir um único valor de raio por classe, foram considerados dois critérios de votação: um primeiro, a votação simples (**VS**), que seleciona o candidato apontado por 2 dentre os 3 índices; e um segundo, a votação ponderada (**VP**), que buscou solucionar eventuais empates do primeiro critério. O critério de votação ponderada foi realizado pelo produto de três curvas, aqui referidas como de contribuição do índice, resultando numa única curva, cujo valor de máximo definiu o raio. Para os índices SH e DUNN, estas curvas foram produzidas pela divisão das curvas de valores do índice por seus respectivos valores máximos. Para o índice DB, cuja seleção é baseada no valor mínimo, realizou-se o inverso dos valores, dividindo-se o resultado pelo maior valor encontrado. Deste modo, para cada valor de raio é associado, na curva de contribuição de um índice particular, um valor entre 0 e 1, que define um peso. A escolha do melhor raio é definida pelo raio que apresentar o maior peso total, determinado através do produto dos pesos atribuídos por cada índice.

Na Tabela 5.5 são resumidos os valores de raio apontados pelos 2 critérios para cada modalidade de compactação. É possível perceber que, para maioria das classes, ambos critérios apontam para o mesmo valor de índice. Para as divergências, há duas situações: uma primeira, onde os índices sinalizam em direção a um valor de raio grande ou pequeno; e uma segunda, onde o critério VP pode ser utilizado como desempate dos candidatos fornecidos pelo critério VS. Para o primeiro caso, tem-se

a classe E2, para 5 componentes de discriminação, e as classes E2 e G para 5 e 10 componentes de representação, respectivamente. Para o segundo, tem-se a classe E2, para 10 componentes de discriminação, e as classes C e H para 5 componentes de representação. No desempate, o primeiro resulta num raio grande (5,0), enquanto os demais, num raio pequeno (0,5). Na Tabela 5.6 é apresentado um resumo dos raios escolhidos pelos índices por classe e modalidade de compactação considerando os critérios VS e VP.

Tabela 5.5: Valores dos raios de vigilância propostos pelos critérios de votação por classe (Veja o texto).

PCD-10C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
VS	0,5	0,5	0,5	3,0	0,5	0,5/4,5/5,0	5,0	5,0	0,5
VP	0,5	0,5	0,5	3,0	0,5	5,0	5,0	5,0	0,5
PCA-10C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
VS	5,0	3,5	5,0	4,5	4,5	5,0	5,0	4,0	5,0
VP	5,0	3,5	5,0	4,5	4,5	5,0	5,0	4,5	5,0
PCD-5C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
VS	5,0	0,5	0,5/3,5/4,5	4,5	0,5	4,0	5,0	3,5	0,5/2,0/5,0
VP	5,0	0,5	0,75	4,5	0,5	5,0	5,0	3,5	0,5
PCA-5C	**A**	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
VS	4,5	3,5	5,0	4,5	4,5	5,0	4,5	4,5	5,0
VP	4,5	3,5	5,0	4,5	4,5	4,5	4,5	4,5	5,0

Tabela 5.6: Valores dos raios de vigilância selecionados (Veja o texto).

	A	**B**	**C**	**D**	**E1**	**E2**	**F**	**G**	**H**
PCD-10C	0,5	0,5	0,5	3,0	0,5	5,0	5,0	5,0	0,5
PCA-10C	5,0	3,5	5,0	4,5	4,5	5,0	5,0	4,0	5,0
PCD-5C	5,0	0,5	0,5	4,5	0,5	4,0	5,0	3,5	0,5
PCA-5C	4,5	3,5	5,0	4,5	4,5	5,0	4,5	4,5	5,0

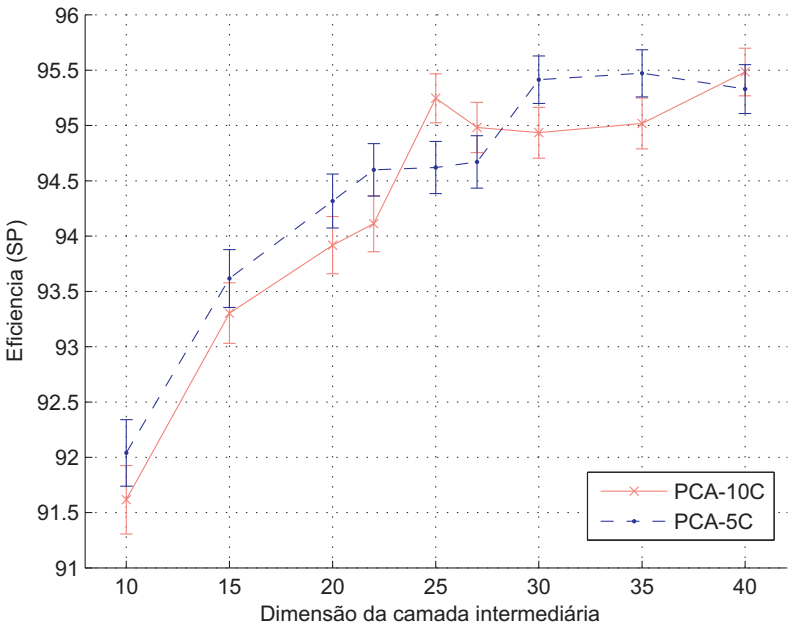
Através da Tabela 5.6, para 10 componentes de discriminação, pode-se observar que um valor de raio pequeno (0,5) foi escolhido para 5 das 9 classes envolvidas (A, B, C, E1 e H); um valor de raio grande (5,0) para as classes E2, F e G, e apenas

para a classe D foi selecionado um valor de raio médio (3,0). Para 5 componentes, em metade das classes (B, C, E1 e H), o raio selecionado foi pequeno (0,5); na outra metade (A, D, E2 e F), um valor de raio grande (entre 4,0 e 5,0) e, por fim, um raio médio (3,5) para a classe G. Considerando 5 ou 10 componentes de discriminação, em 6 das 8 classes há forte coincidência quanto às indicações de raio: pequeno para as classes B, C, E1 e H, e grande para as classes E2 e F. Para 5 ou 10 componentes de representação, há uma indicação de raio grande (entre 4,0 e 5,0) para todas as classes, exceto a classe B, cujo raio indicado é médio (3,5).

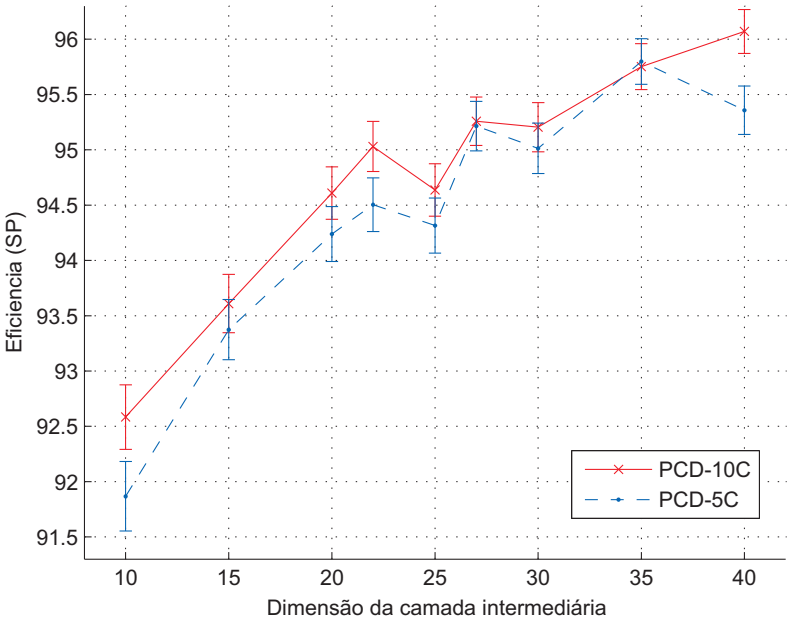
Dos resultados anteriores, é interessante observar que os índices indicaram valores extremos de raio (pequeno e grande) para a maior parte das classes, o que sinaliza em favor de duas granularidades para o tratamento das classes: uma mais fina e outra mais grosseira. Considerando as componentes de discriminação, para metade das classes é indicada uma granularidade fina. Para as componentes de representação, para quase todas as classes é indicada uma granularidade grosseira. Novamente, há um indicativo que as componentes de representação produzem dados com uma estrutura menos complexa, visto que o agrupamento identificado pelos índices como aquele que melhor reflete a estrutura existente nos dados é o de granularidade mais grosseira, logo com um menor número de grupos.

Para identificar qual compactação apresenta melhor desempenho, foi produzido um classificador por modalidade de compactação, utilizando conjuntos de projeto e avaliação baseados nos agrupamentos selecionados pelos índices. Na Figura 5.18 são apresentadas as eficiências (SP) correspondentes aos classificadores produzidos com base nas 4 modalidades de compactação. Em relação as componentes de representação, há um melhor desempenho para a compactação em 5 componentes, a qual apresenta uma vantagem de $0,3$ (40 neurônios) a $0,7$ (25 neurônios) pontos percentuais, exceto para as redes com 25, 27 e 40 neurônios, cuja compactação em 10 compontentes possui uma vantagem de $\approx 0,7$, $\approx 0,4$ e $\approx 0,2$, respectivamente. Considerando as componentes de compactação, o desempenho de 10 componentes é equivalente ou superior a 5 componentes para toda faixa de neurônios, com uma vantagem máxima de $0,7$ (40 neurônios).

Na Figura 5.19 são reunidas as compactações de melhor desempenho para as componentes de representação e discriminação. Verifica-se um melhor desempe-



(a)



(b)

Figura 5.18: Eficiências de generalização (SP) dos classificadores produzidos com base em agrupamentos seqüenciais de espectros que utilizaram uma compactação baseada em componentes de representação e discriminação, com a seleção dos valores de raio vigilância das classes realizada por índices (Veja o texto).

nho da compactação para componentes de discriminação, com vantagem de 0,2 (27 neurônios) a 1,0 (10 neurônios) pontos percentuais, exceto para 25 neurônios, cuja vantagem das componentes de representação é de $0,6$.

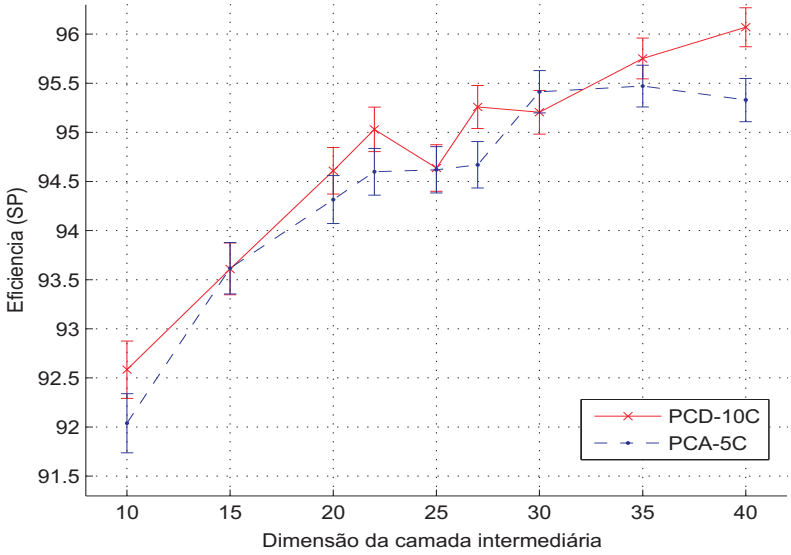


Figura 5.19: Eficiências dos classificadores de melhor desempenho para as compactações em componentes de representação e discriminação, com a seleção dos valores de raio de vigilância das classes realizada por índices (Veja o texto).

É interessante observar que os índices sinalizaram, para maioria das classes, valores de granularidade extremas (pequeno ou grande), as quais foram considerados no processo de seleção por nível arbitrário de granularidade (avaliado na seção anterior). Para ambas propostas de seleção, o conjunto de compactações de melhor desempenho foi idêntico: PCA-5C e PCD-10C, quando analisadas, em separado, as compactações em componentes de representação e discriminação. Considerando todas as modalidades de compactação, para ambas seleções, o melhor desempenho foi obtido pela a PCD-10C.

Na proposta de seleção por nível arbitrário de granularidade, um mesmo valor de raio (0,5 - pequeno) foi considerado para todas as classes. Pelos índices, considerando 10 componentes de discriminação, um valor é proposto para cada classe. Qual destas opções resulta em melhor desempenho? Na Figura 5.20 são apresentadas as

eficiências SP de cada proposta. É interessante observar que a seleção com mesmo nível de granularidade para todas as classes produz classificadores mais eficientes que a baseada nos índices, em torno de $0,1$ (35 neurônios) a $0,9$ pontos percentuais, exceto para 20 neurônios. Este resultado indica que, para o problema em estudo, considerando uma compactação em componentes de discriminação, os agrupamentos seqüenciais baseados numa mesma granularidade (fina) para todas as classes melhor refletem as características dos dados relevantes à seleção dos conjuntos.

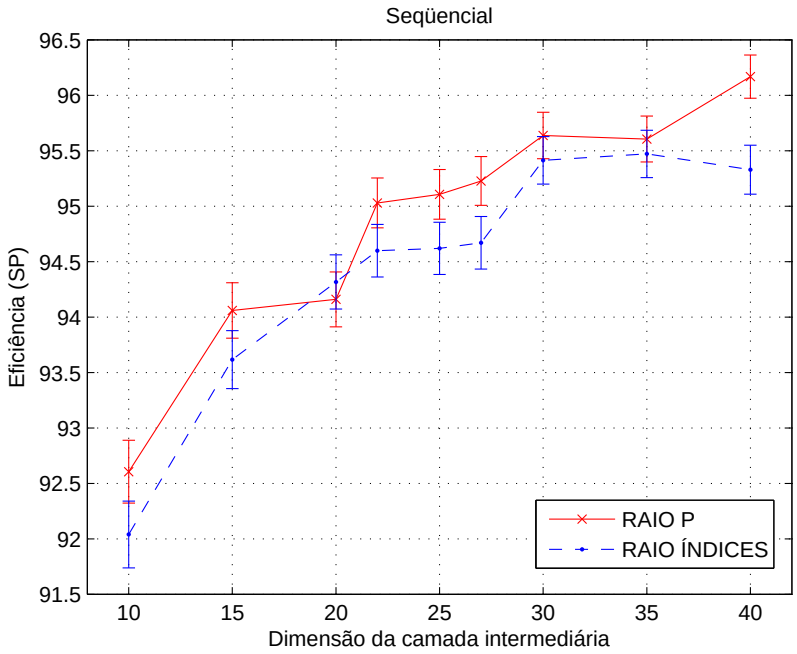


Figura 5.20: Comparação das eficiências de generalização (SP) dos classificadores produzidos com base no agrupamento seqüencial dos espectros, considerando a seleção dos raios através do critério de arbitrário de granularidade (raio P) e através da utilização dos índices DB, DUNN e Silhouette (raio índices).

5.1.5.3 Agrupamento hierárquico

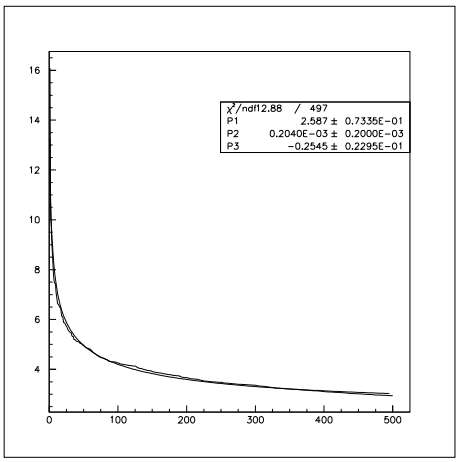
Para a produção dos agrupamentos hierárquicos, a distância euclidiana e o critério de ligação média foram considerados como medida de dissimilaridade entre eventos e grupos, o último escolhido por sua maior imunidade a eventos espúrios. Para a compactação dos dados foram utilizadas 10 componentes de discriminação,

e uma mesma granularidade arbitrária foi considerada para o corte dos dendrogramas de todas as classes, escolhas que foram baseadas nos resultados obtidos para o agrupamento seqüencial (vide seção 5.1.5.2).

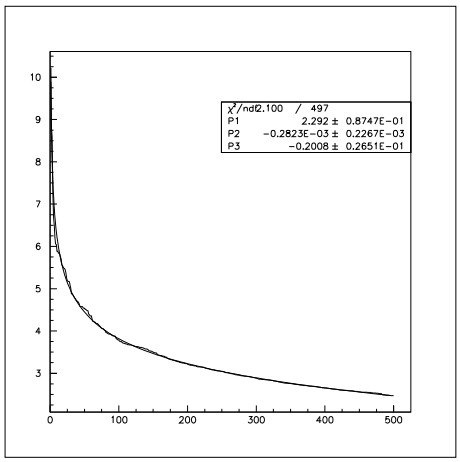
De posse dos agrupamentos de cada classe, foi produzida a curva de dissimilaridade segundo procedimento detalhado na seção 5.1.3.1, sendo realizado o ajuste da curva definida pela função da Equação 5.5 através do software estatístico PAW (*Physics Analysis Workstation*) [182]. Nas Figuras 5.21 e 5.22 são apresentados, classe-a-classe, as curvas original e ajustada, os valores das constantes (p_1,p_2 e p_3) e suas incertezas, assim como o valor do qui-quadrado (χ^2) [52] do processo de ajuste, que é uma medida do nível de confiabilidade do modelo proposto. É possível perceber que a função proposta realiza um bom ajuste das curvas originais, o que é confirmado pelo valor da estatística χ^2 em relação ao número de graus de liberdade (*ndf -number of degrees of freedom*). As classes de ajuste mais crítico foram a D e a G, as quais apresentaram curvas de decaimento mais acentuado.

Com base nos parâmetros obtidos pelo processo de ajuste das curvas, foram elaboradas curvas relacionando o número de grupos produzidos pelo corte dos dendrogramas, segundo diferentes números de constantes de decaimento, para todas as classes. Estas curvas são exibidas na Figura 5.23. É possível perceber que os grupos de classes (A, E1 e E2) e (F e H) possuem curvas similares, logo estão associados a curvas de dissimilaridade com decaimento similar. As classes D e G apresentam um pequeno número de grupos para a faixa considerada, dada a forma abrupta exibida por suas curvas de dissimilaridade. Observa-se uma tendência de crescimento exponencial do número de grupos a medida que um maior número de constantes de decaimento é considerado. Através da análise destas curvas, foram arbitrados, subjetivamente, às granularidades grosseira, intermediária e fina, os seguintes números de constantes de decaimento: 0,5, 0,9 e 1,2, aqui referidos como pequeno (P), médio (M) e grande (G). Segundo estas constantes, para a classe B, que está associada à curva de crescimento mais abrupto, tem-se agrupamentos com 12, 79 e 270 grupos. Para a classe A, cuja curva possui um crescimento moderado, tem-se 7, 35 ou 123 grupos. As classes D e G, cujas curvas apresentam um pequeno crescimento, tem-se 2, 3 ou 4 grupos.

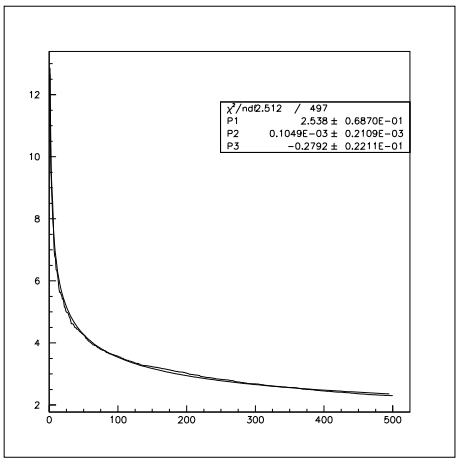
A identifição da granularidade mais adequada seguiu procedimento análogo



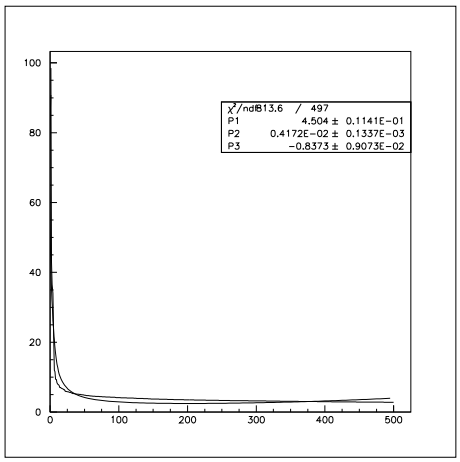
(a) Classe A



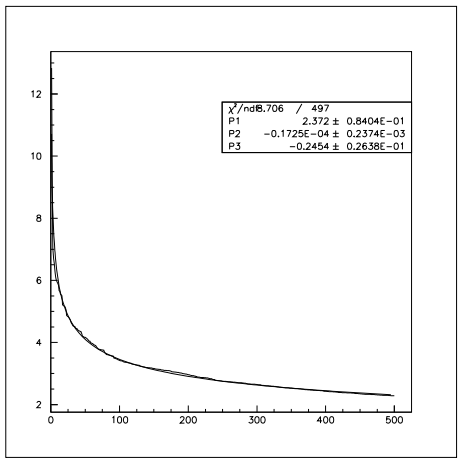
(b) Classe B



(c) Classe C

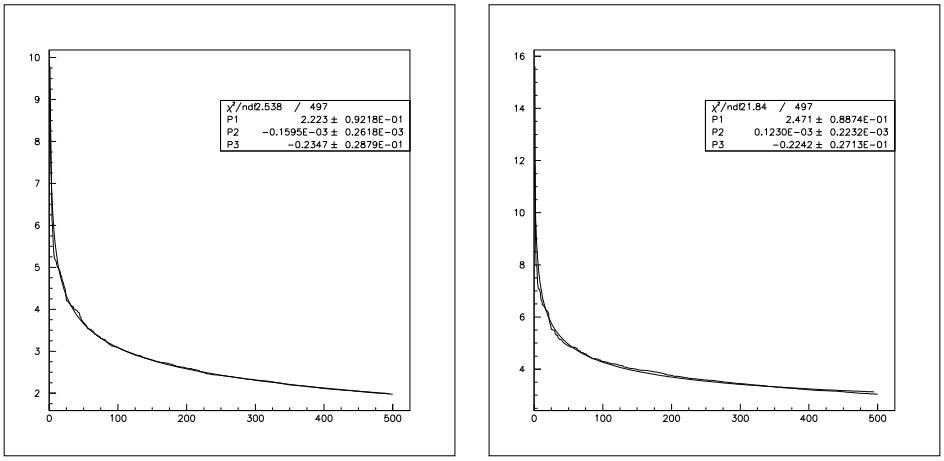


(d) Classe D

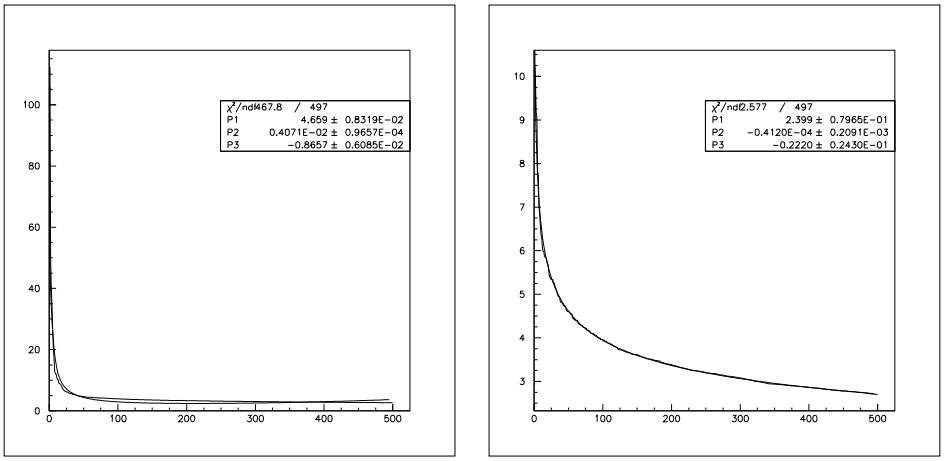


(e) Classe E1

Figura 5.21: Ajuste do modelo proposto para o agrupamento hierárquico baseado em 10 componentes de discriminação (classes A a E1)



(a) Classe E2 (b) Classe F



(c) Classe G (d) Classe H

Figura 5.22: Ajuste de modelo para agrupamento hierárquico baseado em 10 componentes de discriminação (classes E2 a H)

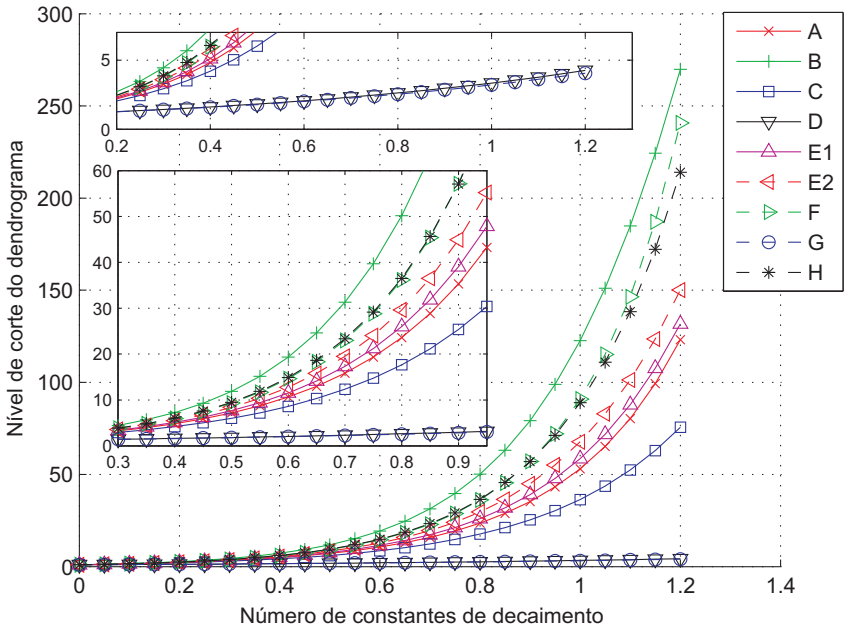


Figura 5.23: Número de grupos do dendrograma por classe e número de constantes de decaimento para o agrupamento hierárquico dos espectros.

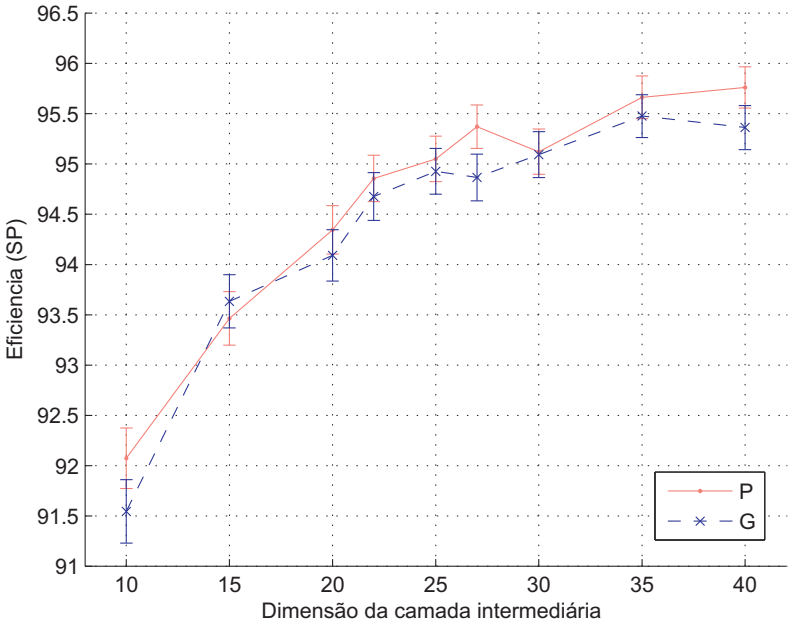
ao realizado para o agrupamento seqüencial. Novamente, por restrições quanto ao custo computacional, uma mesma granularidade foi utilizada para o corte de todas as classes. Os valores de eficiência de generalização (SP), por escolha da constante de decaimento (P, M e G) e faixa de neurônios na camada intermediária dos classificadores são apresentados na Figura 5.24. É possível perceber que a granularidade pequena (P) tende a apresentar resultados de eficiência iguais ou superiores as demais granularidades. Na Figura 5.24(a), na comparação entre as constantes P e G, a primeira resulta em eficiências superiores de ≈ 0 (30 neurônios) a $\approx 0,6$ (27 neurônios) pontos percentuais, exceto para 15 neurônios. Na comparação entre P e M (Figura 5.24(a)), as diferenças são menores, de ≈ 0 (30 neurônios) a $\approx 0,4$ (27 neurônios), exceto para 25 e 40 neurônios.

5.1.5.4 Comparação das modalidades de seleção

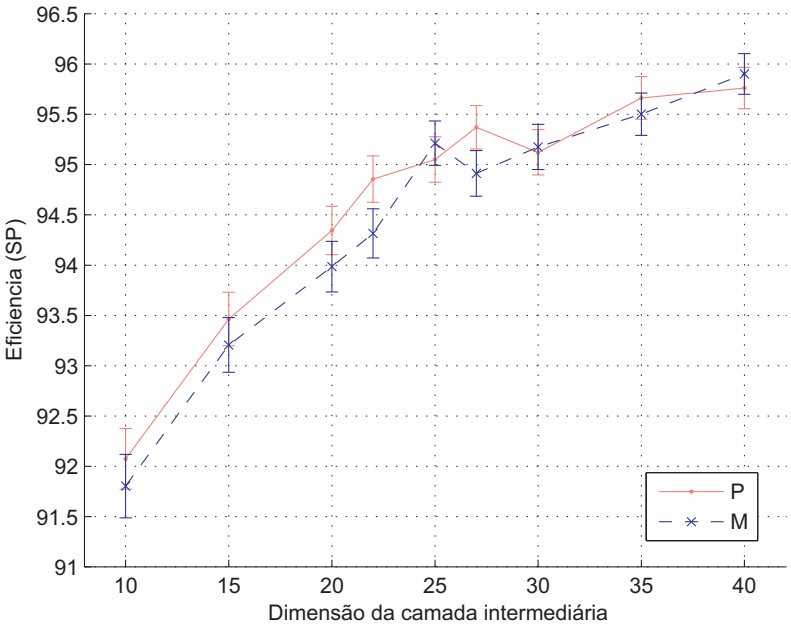
A identificação dos melhores parâmetros para os agrupamentos seqüencial e hierárquico motiva as seguintes perguntas: qual destes agrupamentos provê uma melhor seleção dos conjuntos de projeto e avaliação? Quais seriam as vantagens de sua utilização em relação ao processo de seleção baseado em subamostragem alea tória, aqui referido como de sorteio, que foi descrito na seção 4.3.

Na Figura 5.25 são apresentadas as curvas de eficiência (SP) obtidas para três modalidades de seleção dos conjuntos: duas baseadas em agrupamentos (seqüencial e hierárquico), e outra, através de sorteio. Quando comparados os agrupamentos seqüencial e hierárquico (Figura 5.25(a)), verifica-se uma vantagem do primeiro de ≈ 0 (25 neurônios) a $\approx 0,5$ (30 neurônios) pontos percentuais, exceto para os classificadores com 20, 27 e 35 neurônios. Na comparação entre o agrupamento seqüencial e o sorteio, as diferenças observadas são superiores, com uma vantagem do primeiro de ≈ 0 (10 neurônios) a $\approx 0,9$ (25 neurônios) pontos percentuais.

Dos resultados, verifica-se que a seleção dos conjuntos por agrupamento seqüencial mostrou-se mais eficaz, tanto em termos de desempenho quanto no custo computacional, o qual é significativamente inferior, sendo bastante atrativa para aplicações com extenso volume de dados, tais como o sonar passivo.

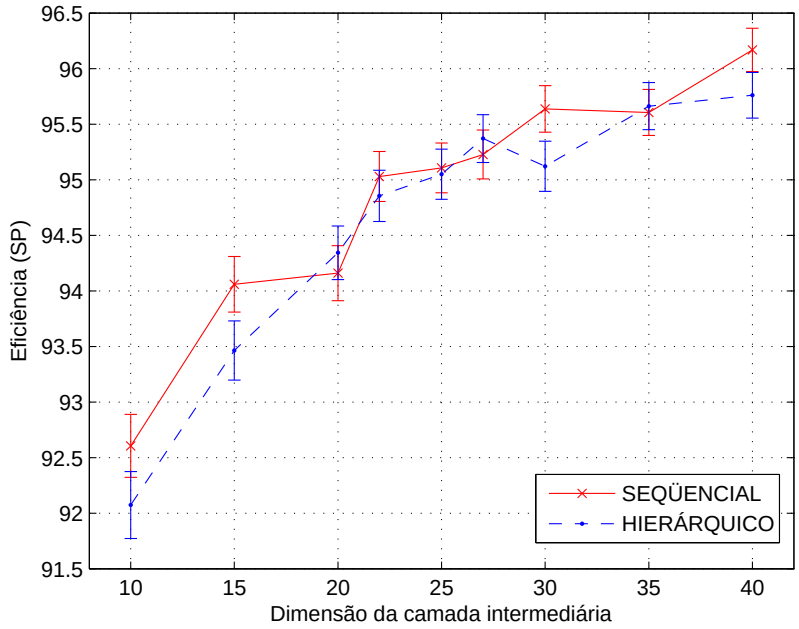


(a)

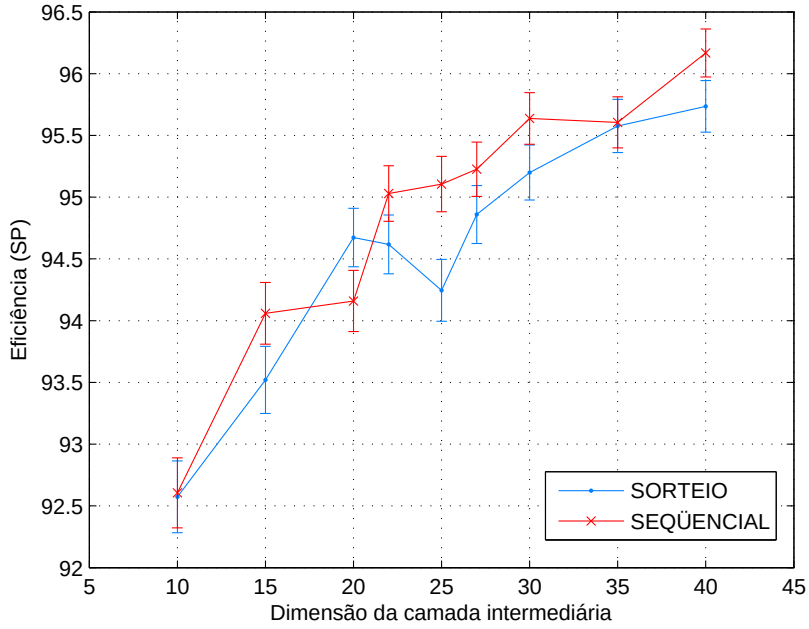


(b)

Figura 5.24: Eficiências de generalização (SP) dos classificadores produzidos através do corte do agrupamento hierárquico dos espectros, para diferentes níveis subjetivos de granularidade, considerando dados compactados em 10 componentes de discriminação (Veja o texto).



(a)



(b)

Figura 5.25: Eficiências de generalização (SP) dos classificadores produzidos através da seleção dos conjuntos por agrupamento seqüencial, hierárquico e por sorteio (Veja o texto).

5.2 Análise de agrupamentos para a seleção por corridas

Conforme discutido anteriormente, a utilização da seleção por corridas resulta na produção e avaliação do sistema de classificação em condições mais próximas a sua operação real, a qual contemplará condições operativas, navios ou mesmo classes desconhecidas. Para proceder esta seleção, de forma que os conjuntos de projeto e avaliação retenham uma estatística representativa do problema, faz-se necessário identificar similaridades entre as diferentes corridas.

Nas seções anteriores foi discutida a identificação de similaridades entre eventos através de técnicas de agrupamento. Com respeito às corridas, agrupamentos também podem ser utilizados, ainda que a identificação de similaridades e a formação dos grupos deva ser baseada, no entanto, em subconjuntos de eventos. Há medidas de similaridade entre grupos de eventos, utilizadas para a produção de agrupamentos hierárquicos (vide seção 5.1.3). Não foram encontradas, no entanto, referências que tratassem de agrupamentos de subconjuntos de dados.

Estes fatores motivaram o desenvolvimento de um critério para a identificação de similaridades entre as corridas. Por este critério, a cada corrida é associado um vetor representativo. De posse dos vetores representativos de cada corrida, produz-se um um agrupamento, sendo identificados quais vetores partilham um mesmo grupo, os quais identificam corridas similares.

Para a produção dos vetores representativos, optou-se por utilizar os agrupamentos seqüenciais, classe-a-classe, aqui referidos como agrupamentos base, visto seu melhor desempenho na seleção baseada em espectros. A geração destes vetores foi baseada na extração de informações relativas à maneira como os eventos de cada corrida estão distribuídos na estrutura identificada por cada agrupamento, segundo critérios a serem discutidos nas seções 5.2.1 e 5.2.2. A Figura 5.26 resume o método proposto.

Inicialmente, serão descritos dois critérios para a produção de vetores representativos das corridas. Em seguida, será discutida a produção dos agrupamentos com base nestes vetores representativos. Por fim, serão apresentados resultados referentes à aplicação das técnicas propostas na seleção dos conjuntos.

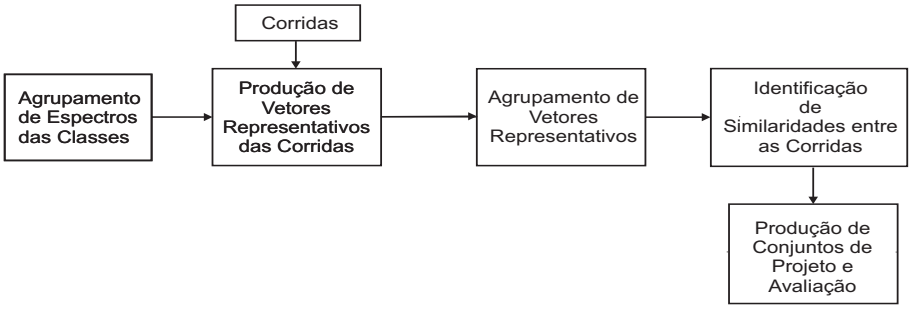


Figura 5.26: Sistemática para a avaliação de similaridades entre as corridas.

5.2.1 Produção de vetores representativos com base na pertinência das corridas aos grupos e nas coordenadas dos seus centros

Nesta proposta são identificados no agrupamento base, para cada corrida, quais os grupos possuem eventos a ela pertencentes. Estes grupos serão aqui referidos como excitados pela corrida. O vetor representativo de cada corrida será definido pelo centro de gravidade (baricentro) ou média dos vetores correspondentes aos centros de gravidade dos grupos por ela excitados.

Para fins ilustrativos, na Figura 5.27 são apresentadas três corridas arbitrárias ($C1$, $C2$ e $C3$), de dados bidimensionais, cujos eventos estão representados pelos símbolos + ($C1$), x ($C2$) e $*$ ($C3$). O agrupamento base (seqüencial) possui três grupos $g1$, $g2$ e $g3$, que são indicados através de círculos. Os centros de gravidades destes grupos, indicados por quadrados, possuem vetores **cg1**, **cg2** e **cg3**.

É possível observar que a corrida $C1$ está distribuída nos três grupos. A corrida $C2$, nos grupos 1 e 3; e a $C3$, nos grupos 2 e 3. Os vetores representativos das corridas (**cc1**, **cc2** e **cc3**), indicados por triângulos, são calculados da seguinte forma:

$$\mathbf{cc1} = \frac{\mathbf{cg1} + \mathbf{cg2} + \mathbf{cg3}}{3} \tag{5.27}$$

$$\mathbf{cc2} = \frac{\mathbf{cg1} + \mathbf{cg3}}{2} \tag{5.28}$$

$$\mathbf{cc3} = \frac{\mathbf{cg2} + \mathbf{cg3}}{2} \tag{5.29}$$

Uma alternativa mais simples ao cálculo proposto pelas Equações 5.27-5.29

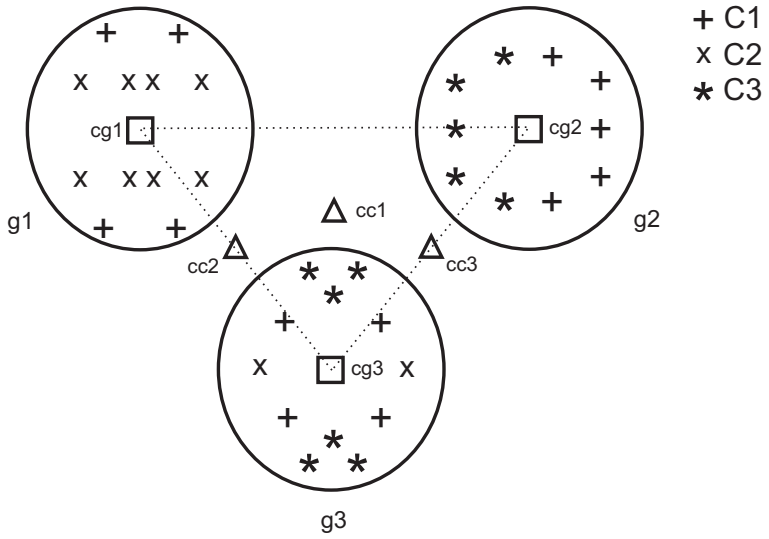


Figura 5.27: Corridas, grupos e vetores representativos definidos através do baricentro dos centros excitados (Veja o texto).

seria determinar os vetores representativos de cada corrida pelo baricentro dos eventos a ela pertencentes, o que não exigiria a utilização de um agrupamento base. Por outro lado, haveria uma tendência destes vetores em situar-se mais próximos de regiões com maiores concentrações de eventos, o que poderia mascarar cenários raros e de interesse. Esta alternativa foi descartada para permitir um melhor balanço entre cenários raros e usuais na formação dos vetores representativos. Outra vantagem da proposta é que o vetor representativo pode também carregar informações de outras corridas, visto que é baseado no centro dos grupos, os quais podem conter eventos de mais de uma corrida.

5.2.2 Produção de vetores representativos com base apenas na pertinência das corridas aos grupos

A idéia principal desta proposta é que corridas similares devam possuir grupos excitados em comum. Neste caso, a identificação de similaridades é baseada na produção de vetores representativos que indicam quais grupos são excitados por cada corrida. Estes vetores serão referidos como vetores de excitação. O agrupamento dos vetores de excitação resulta na formação de grupos que refletem corridas similares,

similaridade que foi aferida através da forma como os eventos das corridas estão distribuídos pelos grupos do agrupamento.

Na produção dos vetores de excitação, para um agrupamento de N grupos, é utilizado um vetor de N componentes, onde cada componente responde por um grupo. Caso a corrida possua um ou mais eventos pertencentes a um dado grupo, sua componente correspondente vale 1; em caso contrário, vale zero. Para o exemplo da Figura 5.27, onde as corridas $C1$, $C2$ e $C3$ possuem eventos nos grupos $\{1, 2, 3\}$, $\{1, 3\}$, $\{2, 3\}$, respectivamente, os vetores representativos das corridas seriam: $\begin{bmatrix}1 & 1 & 1\end{bmatrix}$, $\begin{bmatrix}1 & 0 & 1\end{bmatrix}$ e $\begin{bmatrix}0 & 1 & 1\end{bmatrix}$.

5.2.3 Agrupamento dos vetores de pertinência e seleção dos conjuntos de projeto e avaliação

Uma vez produzidos os vetores de pertinência de cada corrida, pelo critério de seleção proposto, há dois cenários possíveis, que são resumidos na Figura 5.28. Numa primeira modalidade, representada na Figura 5.28(a), cada classe é tratada de forma independente, logo a formação dos conjuntos explora apenas as relações existentes entre corridas de uma mesma classe. A segunda modalidade, apresentada na Figura 5.28(b), realiza o agrupamento com base nos vetores representativos de todas as classes, o que resulta em conjuntos que exploram relações entre diferentes classes, o que pode ser útil na identificação de corridas de classificação mais crítica.

O agrupamento de vetores representativos que considere relações entre diferentes classes demanda a produção de um agrupamento base envolvendo todas as classes. Para o conjunto de sonar, em virtude do elevado número de eventos existentes (29277), a produção de um único agrupamento é contra-indicada pelo custo computacional, razão pela qual todos agrupamentos considerados neste trabalho foram produzidos classe-a-classe. Para vetores representativos baseados no centro dos grupos excitados, é possível, no entanto, derivar um critério alternativo que, mesmo baseado em agrupamentos classe-a-classe, avalia similaridades entre as diferentes classes. Por este critério, uma mesma normalização e pré-processamento são aplicadas a todas as classes, e os vetores representativos são identificados para cada classe, separadamente. Produz-se um único agrupamento com base nos vetores representativos de todas as classes, o qual permite identificar similaridades entre as

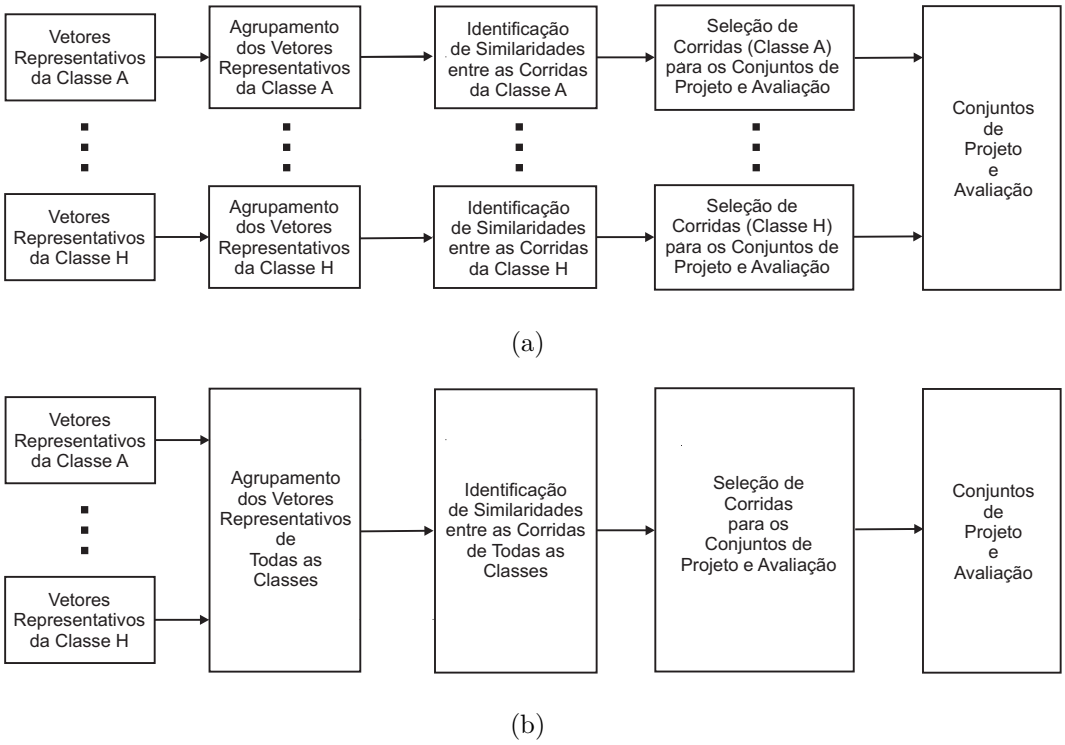


Figura 5.28: Modalidades de agrupamento dos vetores representativos das corridas: (a) classe-a-classe e (b) todas as classes conjuntamente (Veja o texto).

diferentes classes. Deste modo, três critérios de produção dos conjuntos de projeto e avaliação serão aqui considerados: dois, classe-a-classe, utilizando vetores representativos pelas propostas das seções 5.2.1 e 5.2.2; e um, para todas as classes, segundo o último critério.

Em virtude de um melhor desempenho do agrupamento seqüencial na seleção dos conjuntos por espectros (conforme a seção 5.1.5.4), esta técnica foi considerada, inicialmente, para a produção dos agrupamentos de vetores representativos. Constatou-se, no entanto, para diferentes execuções do algoritmo, uma flutuação expressiva no número e no conteúdo dos grupos, a qual mostrou-se relacionada à dependência do agrupamento seqüencial com respeito a ordem de apresentação dos eventos, aleatória a cada execução. Tanto menor o número de vetores representativos disponíveis, entre 22 (classe A) e 43 (classe C) [6], para o agrupamento classe-a-classe, e 263, para o agrupamento com todas as classes, mais acentuadas foram as flutuações observadas. Em virtude desta deficiência, e dado que nenhuma técnica foi encontrada na literatura para contornar este problema, optou-se pela utilização da técnica hierárquica, explorando o critério de corte do dendograma proposto na seção 5.1.3.1.

Para a produção do agrupamento hierárquico, quando considerados vetores representativos baseados nos centros do grupos excitados, foi considerada a distância euclidiana para a medida de similaridade entre os eventos, e o critério de ligação média para a formação dos grupos, ambos discutidos na seção 5.1.3. Para vetores representativos definidos por vetores de excitação, no entanto, a distância euclidiana não é a mais indicada, visto que estes vetores são binários.

Algumas alternativas da literatura para medir a similaridade entre vetores binários são os índices de Rao [183], Jaccard [172], Rand [175], Rand Ajustado [161], Dice [184], entre outros. Segundo estes índices, a comparação de dois vetores binários $\mathbf{V}_1$ e $\mathbf{V}_2$ é baseda no número de coordenadas onde: (a) ambos vetores valem um; (b) $\mathbf{V}_1$ possui coordenada UM e $\mathbf{V}_2$ ZERO; (c) o contrário de (b) e (d) o contrário de (a).

Para o problema de sonar, segundo o critério da seção 5.2.2, a similaridade

[6] Vale notar que a classe E, com 66 corridas, foi dividida nas classes E1 e E2, com 33 corridas cada uma.

entre duas corridas é medida pelo número de grupos que ambas possuem em comum. Assim, nesta medida de similaridade é importante considerar a quantidade de coordenadas em que ambos vetores valem um, assim como aquelas onde um dos vetores possui valor um e o outro zero, ou seja, em que uma corrida excita um grupo que não é excitado pela outra. O índice escolhido, visto que atende a estes objetivos, é o de Jaccard [172], definido como:

$$I = \frac{n_a}{n_a + n_b + n_c}, \tag{5.30}$$

onde n_a, n_b, n_c e n_d são o número de eventos identificados nas situações (a),(b),(c) e (d), respectivamente. Como critério de similaridade entre grupos também foi utilizado o critério de ligação média.

De posse dos grupos fornecidos pelo agrupamento, a seleção dos conjuntos considerou uma divisão meio-a-meio e aleatória das corridas, indicadas por cada grupo, entre os conjuntos de projeto e avaliação, visto que são tidas como similares. Caso os grupos apresentassem um número ímpar $(2n + 1)$ de vetores representativos, $n + 1$ corridas foram destinadas ao conjunto de projeto; e n, ao conjunto de avaliação, de forma que o aprendizado do classificador fosse privilegiado. Os grupos com um único vetor representativo tiveram suas corridas particionadas, também aleatoriamente e meio-a-meio, entre ambos conjuntos, visto que não havia nenhuma informação ou conhecimento a priori sobre os dados que justificasse sua alocação a um conjunto particular.

5.2.4 Resultados para a seleção estatística baseada em corridas

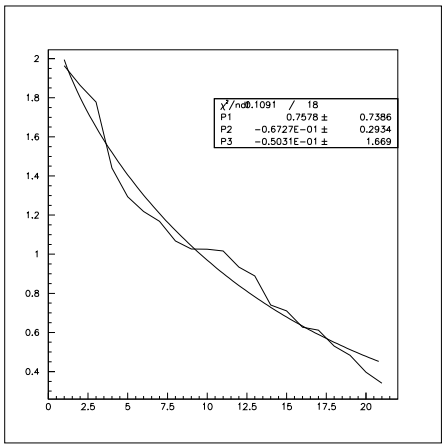
A seguir serão apresentados os resultados referentes à cada modalidade de seleção proposta. Em seguida, as diferentes modalidades serão comparadas com a seleção baseada em sorteio (vide seção 4.3), para identificar qual é a mais apropriada para a seleção baseada em corridas.

5.2.4.1 Seleção baseada em vetores representativos definidos pela pertinência das corridas aos grupos e nas coordenadas dos seus centros (classe-a-classe)

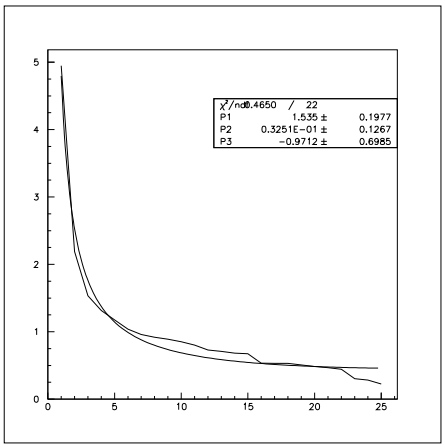
As curvas de dissimilaridade e seus modelos matemáticos obtidos para esta modalidade são esboçados nas Figuras 5.29 e 5.30. É possível observar que as curvas de todas as classes foram modeladas de forma bastante satisfatória. Comparando esta modelagem com a realizada na seleção por espectros, verifica-se uma maior divergência entre as curvas ajustada e original, o que está relacionado às restrições estatísticas relacionadas ao pequeno número de vetores representativos disponível para cada classe.

As curvas do número de grupos do agrupamento para cortes do dendrograma segundo diferentes frações de constantes de decaimento são apresentadas na Figura 5.31. É possivel observar que as classes A, F, E1 e E2 apresentam um crescimento mais acentuado, o que está associado ao decaimento mais expressivo de suas curvas de dissimilaridade. Os pares de classes { C, H } e { B, D } apresentaram curvas bastante similares. Através da inspeção destas curvas, foram atribuídos aos valores subjetivos de granularidade grosseira, intermediária e fina as constantes de decaimento 0,4, 0,9 e 1,4, aqui referidas como pequena (P), média (M) e grande (G). Para a classe A, cuja curva exibe o maior crescimento, este conjunto de constantes resulta em agrupamentos com 6, 13 e 20 grupos, respectivamente. Para a classe E1, de crescimento intermediário, tem-se 3, 7 e 13 grupos. A classe G, de menor crescimento, exibe 2, 3 ou 4 grupos.

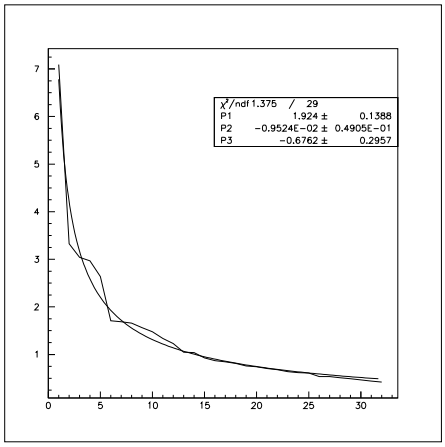
De forma similar ao agrupamento hierárquico de espectros (vide seção 5.1.5.3), inicialmente, um mesmo valor de granularidade foi considerado para o corte de todas as classes. As eficiências de generalização SP associadas aos classificadores produzidos para cada granularidade são apresentadas na Figura 5.32, considerando topologias com 10 a 50 neurônios na camada intermediária. Para quase todos os classificadores, verifica-se um melhor desempenho associado à constante de decaimento pequena, logo a uma granularidade mais grosseira, resultado similar ao obtido na seleção por espectros (vide seção a 5.1.5.3). Na comparação P e G, tem-se um melhor desempenho da primeira de 0 (25 ou 35 neurônios) a $\approx 1,4$ (10 neurônios) pontos percentuais, exceto para redes com 30 neurônios. Para as constantes P e M,



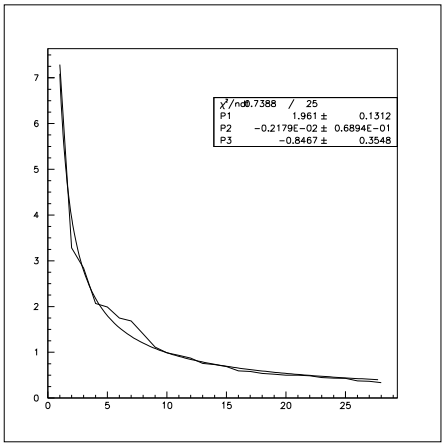
(a) Classe A



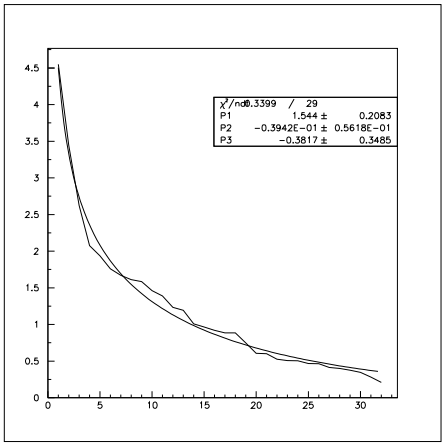
(b) Classe B



(c) Classe C

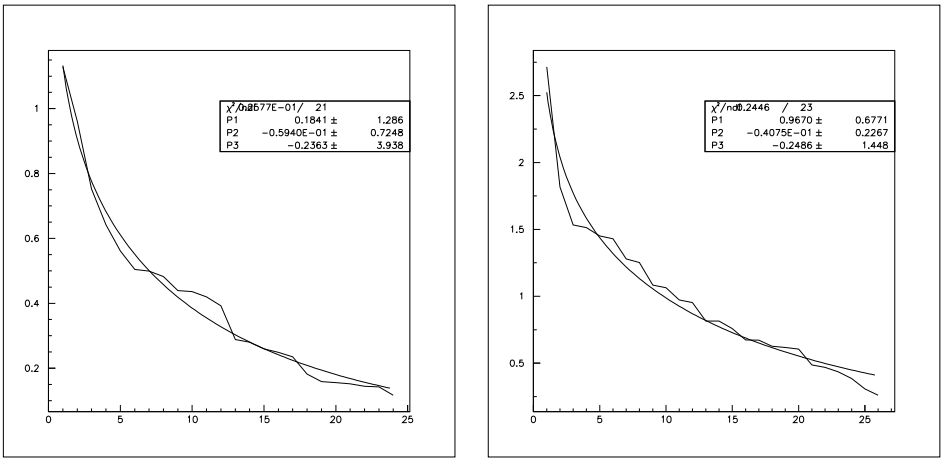


(d) Classe D

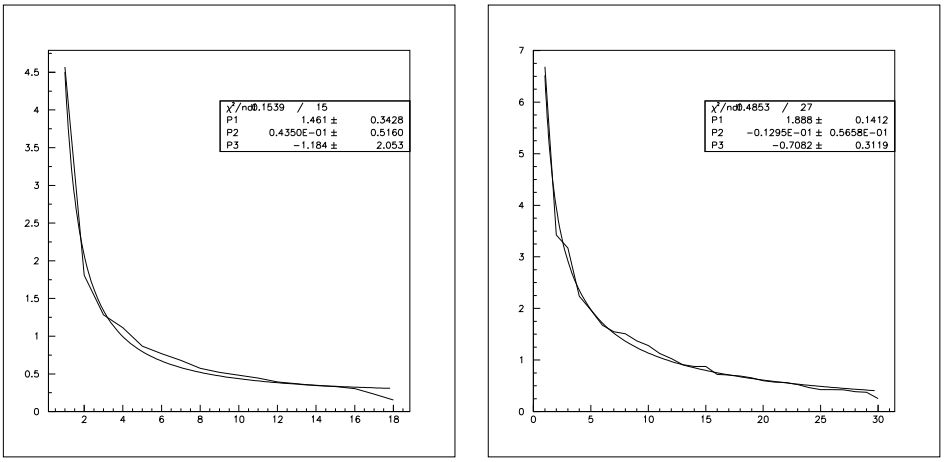


(e) Classe E1

Figura 5.29: Modelagem da curva de dissimilaridade do agrupamento hierárquico, classe-a-classe, para vetores representativos definidos pelo baricentro dos centros excitados pelas corridas (classes A a E1).



(a) Classe E2 (b) Classe F



(c) Classe G (d) Classe H

Figura 5.30: Modelagem da curva de dissimilaridade do agrupamento hierárquico, classe-a-classe, para vetores representativos definidos pelo baricentro dos centros excitados pelas corridas (classes E2 a H).

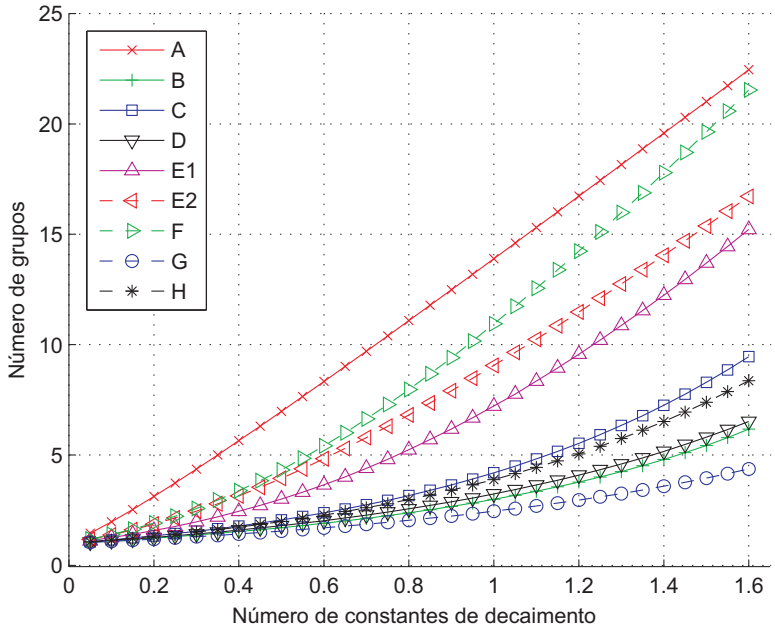


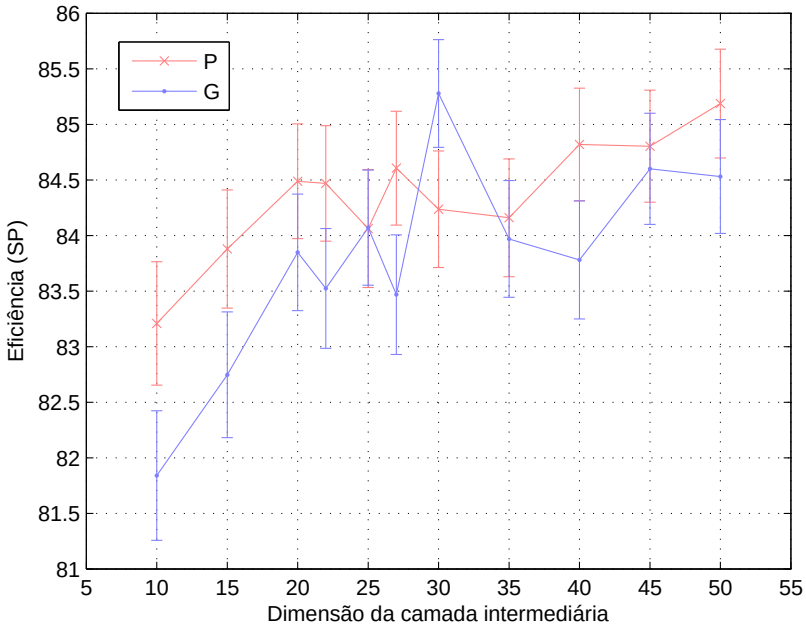
Figura 5.31: Número de grupos do dendrograma por classe e número de constantes de decaimento para o agrupamento hierárquico de vetores representativos definidos pelo baricentro dos centros excitados pelas corridas (Veja o texto).

as diferenças são maiores, de 0 (30 neurônios) a $\approx 2,2$ (10 neurônios). As eficiências associadas à cada constante de decaimento são inferiores àquelas obtidas para a seleção baseada em espectros, visto que a seleção por corridas impõe condições mais severas quanto ao aprendizado e ao teste de generalização dos classificadores.

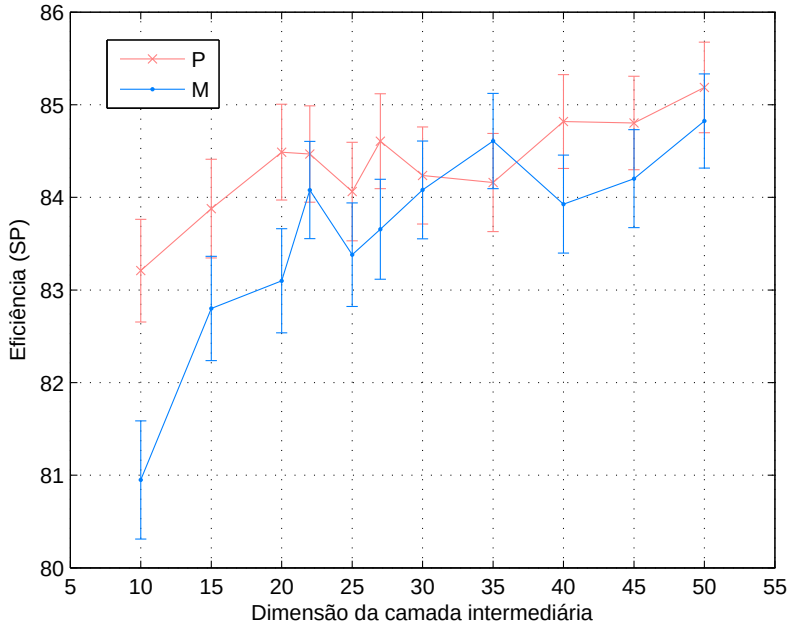
Na simulação anterior, uma mesma granularidade foi considerada para todas as classes. Algumas perguntas seriam: como as eficiências de detecção de cada classe se comportaram com respeito à escolha da granularidade? diferentes granularidades por classe resultariam na produção de um classificador mais eficiente? como selecionar este conjunto de granularidades?

Na Figura 5.33 são exibidas as eficiências de deteção para cada classe e granularidade. Cada classe apresenta um valor de granularidade de melhor desempenho: pequena, para as classes A, C e E; média, para as classes G e H, e grande, para as classes B, D, e F, o que sinaliza em favor da utilização de diferentes granularidades por classe. A seleção de granularidades classe-a-classe demanda testes considerando diferentes combinações, logo possui custo computacionalmente proibitivo. Caso seja pressuposto, no entanto, que a definição da granularidade de uma classe afeta apenas sua eficiência de detecção, um possível critério é identificar, por classe, qual é a granularidade de melhor desempenho. De posse deste conjunto de granularidades, um novo classificador é treinado e avaliado. Se o desempenho deste classificador for superior aqueles produzidos segundo um mesmo corte para todas as classes, confirma-se a hipótese inicial, dispondo-se de um conjunto sub-ótimo de granularidades.

Na Figura 5.34 são comparadas as eficiências de generalização (SP) obtidas para os agrupamentos com granularidades distintas por classe (C), identificadas pelo processo proposto, e segundo uma granularidade pequena (P) para todas as classes. Para as topologias de 10 a 25 neurônios, houve um melhor desempenho da granularidade P, com uma vantagem de 0 (25 neurônios) a $0,6$ (10 neurônios) pontos percentuais. Na faixa de 27 a 50 neurônios, que contempla 6 das 11 redes consideradas, a granularidade C apresentou um desempenho superior, em torno de 0 (27 neurônios) a $0,9$ (35 neurônios). Tendo em vista o número de ensaios e as diferenças de desempenho apresentadas, há um melhor desempenho ao utilizar granularidades distintas.

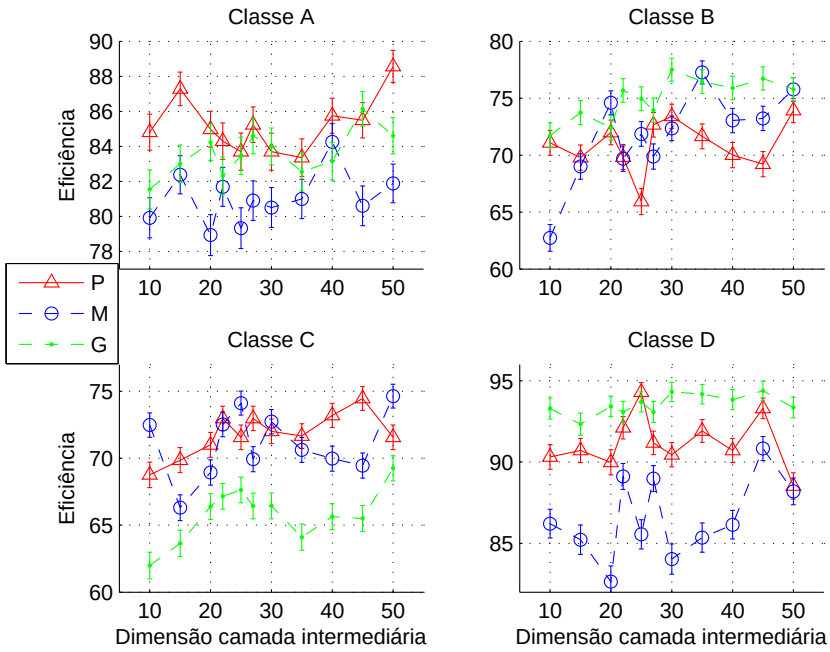


(a)

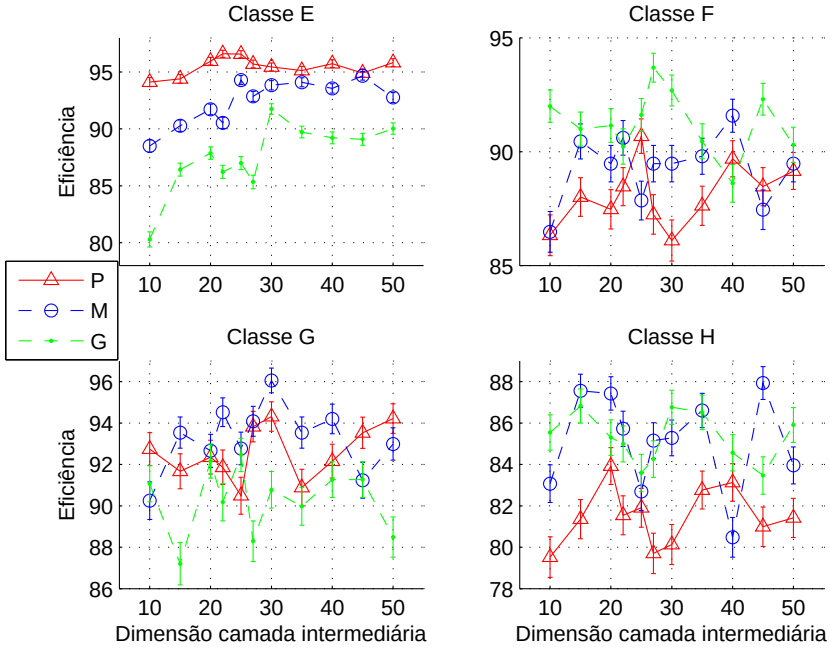


(b)

Figura 5.32: Eficiências de generalização (SP) dos classificadores produzidos com base no corte do agrupamento hierárquico de vetores representativos definidos pelo baricentro dos centros excitados, classe-a-classe, segundo um mesmo nível de granularidade (Veja o texto).



(a)



(b)

Figura 5.33: Eficiências de generalização por classe e nível de granularidade, para uma seleção baseada em corridas, utilizando o agrupamento hierárquico, classe-a-classe, dos baricentros dos centros excitados pelas corridas (Veja o texto).

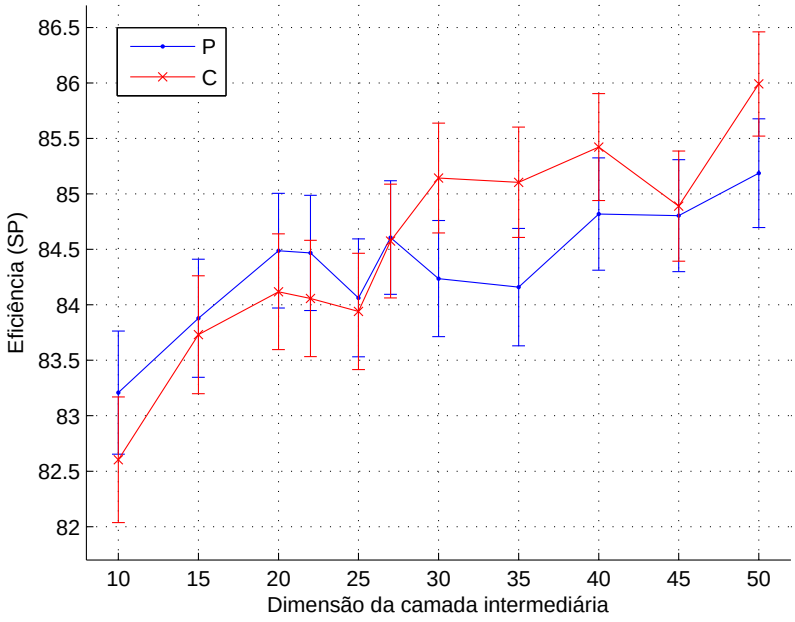
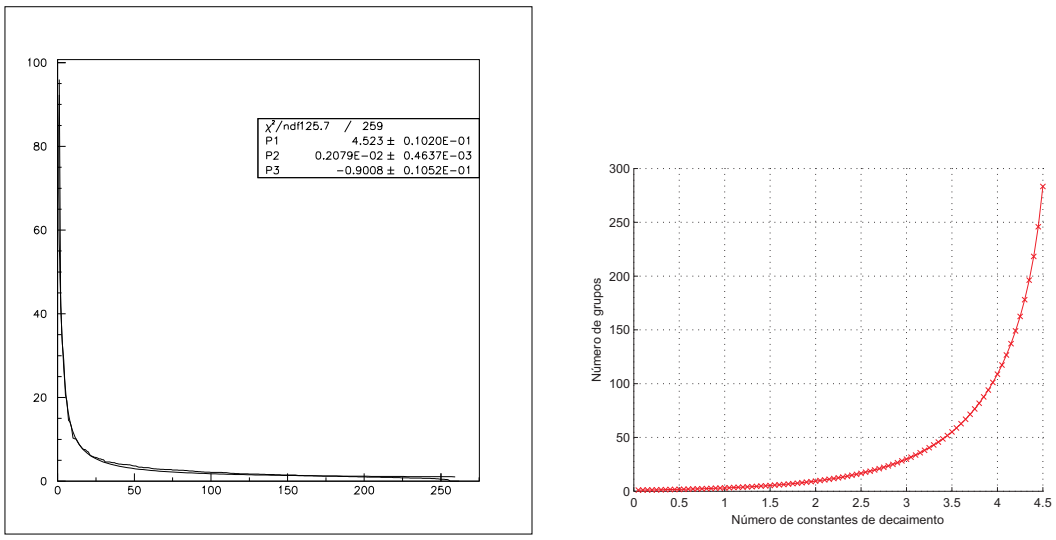


Figura 5.34: Eficiências de generalização dos classificadores obtidos através da seleção de corridas segundo uma mesma (P) ou diferentes granularidades para o corte das classes (C), considerando agrupamentos hierárquicos classe-a-classe dos baricentros dos centros excitados pelas corridas (Veja o texto).

5.2.4.2 Seleção baseada em vetores representativos definidos pela pertinência das corridas aos grupos e nas coordenadas dos seus centros (todas as classes)

Esta proposta se diferencia da apresentada na seção anterior pelo agrupamento dos vetores representativos considerar todas as classes. A curva de dissimilaridade, sua modelagem matemática e a curva do número de grupos por constantes de decaimento são esboçadas na Figura 5.35. Novamente, verifica-se uma modelagem bastante satisfatória da curva de dissimilaridade do agrupamento. Para as granularidades grosseira, intermediária e fina foram atribuídas as constantes de decaimento 3,0, 3,5 e 4,0, referidas como pequena (P), média (M) e grande (G), o que resultou na seleção de agrupamentos com 30, 55 e 109 grupos, respectivamente.

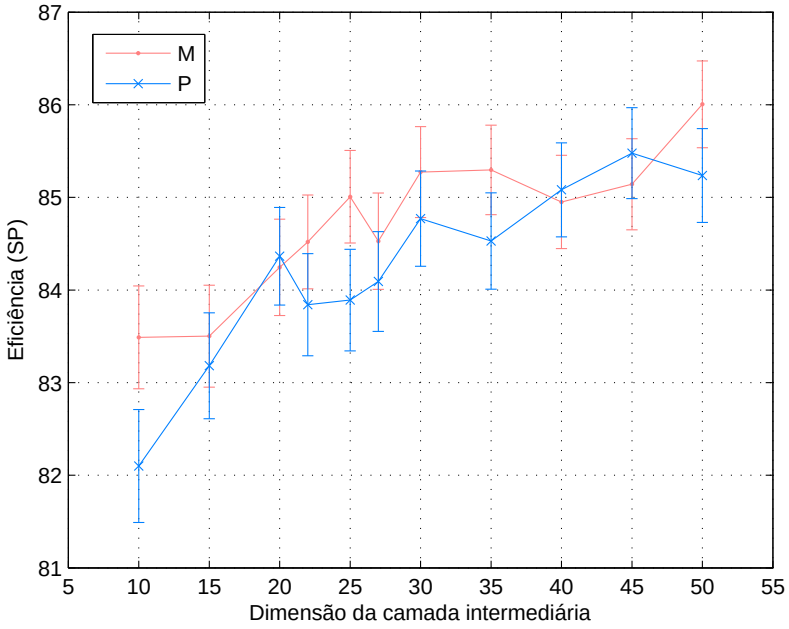


(a) (b)

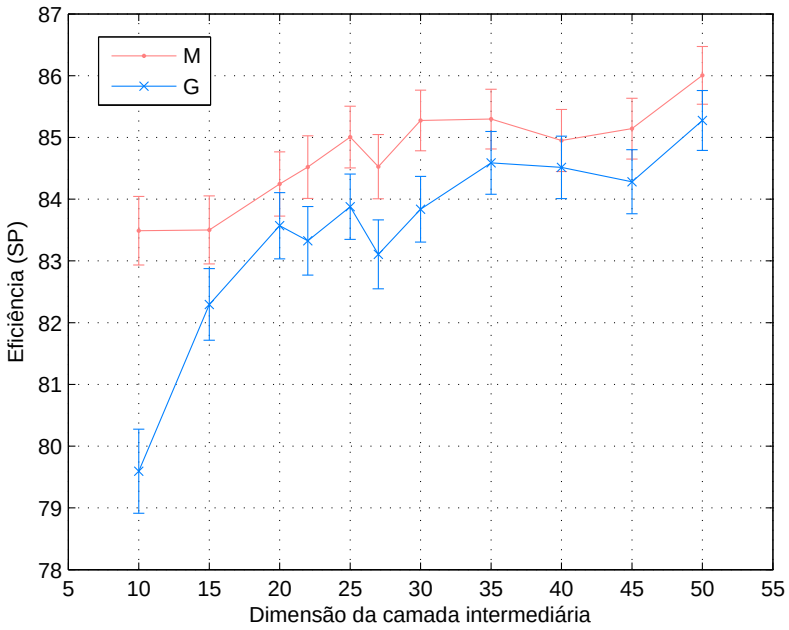
Figura 5.35: (a) Curva de dissimilaridade e modelo matemático para o agrupamento hierárquico dos baricentros dos centros excitados pelas corridas considerando todas as classes e (b) curva do número de grupos do agrupamento por número de constantes de decaimento para o mesmo agrupamento (Veja o texto).

Na Figura 5.36 são apresentadas as eficiências de generalização (SP) dos classificadores segundo diferentes níveis de granularidade. Na comparação entre as constantes de decaimento M e P (Figura 5.36(a)), a primeira resultou numa eficiência de $\approx 0,3$ (15 neurônios) a $\approx 1,4$ pontos percentuais (10 neurônios), exceto para as redes com 20, 40 e 45 neurônios. Para as constantes M e G, o desempenho da

primeira é melhor de $\approx 0,4$ (40 neurônios) a $\approx 3,9$ (10 neurônios) pontos percentuais.



(a)



(b)

Figura 5.36: Eficiências de generalização (SP) dos classificadores produzidos com base no corte do agrupamento hierárquico de vetores representativos definidos pelo baricentro dos centros excitados para todas as classes (Veja o texto).

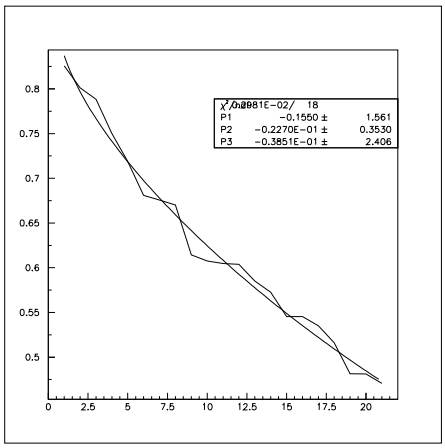
5.2.4.3 Seleção baseada em vetores representativos definidos apenas pela pertinência das corridas aos grupos (classe-a-classe)

Na Figura 5.37 e 5.38 são exibidos, classe-a-classe, as curvas de dissimilaridade e os modelos matemáticos produzidos com base no agrupamento dos vetores de pertinência das corridas ao agrupamento de referência. Considerando as restrições estatísticas existentes, verifica-se um ajuste apropriado dos modelos às curvas de dissimilaridade.

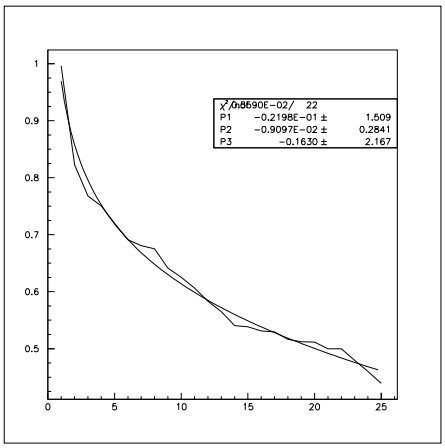
As curvas do número de grupos do agrupamentos por constantes de decaimento, para cada classe, são exibidas na Figura 5.39. Através da análise subjetiva destas curvas, foram atribuídas as constantes de decaimento 0,1, 0,25 e 0,4 às granularidades grosseira, intermediária e fina, respectivamente. Estas curvas apresentam um crescimento mais rápido que as curvas associadas ao agrupamento classe-a-classe dos baricentros excitados (Figura 5.31), o que resultou em menores valores para o conjunto de constantes de decaimento consideradas.

Na Figura 5.40 são apresentadas as eficiências (SP) dos classificadores obtidos para um mesmo nível de granularidade por classe. Quando comparadas as constantes de decaimento P e G, é nítido o melhor desempenho da primeira, com eficiências superiores de $\approx 0,3$ (20 neurônios) a $\approx 2,6$ pontos percentuais (50 neurônios). Na comparação entre P e M, o desempenho é praticamente equivalente, ligeiramente superior para a constante pequena, tendo em vista o desempenho das redes com 10, 15, 22 e 40 neurônios.

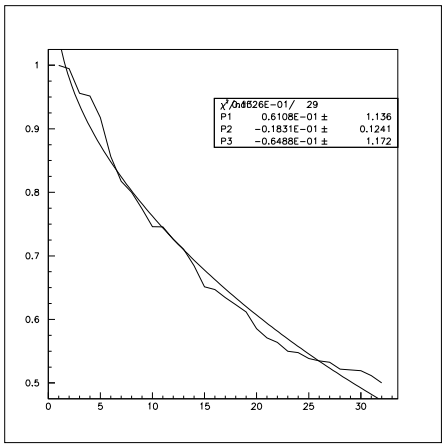
Na Figura 5.41 são apresentadas as eficiências de detecção por classe e granularidade. Para as classes C, E, F e H, há um melhor desempenho da constante de decaimento pequena. Para as classes B e D, da média, e para a A e G, da grande. A Figura 5.42 compara classificadores que consideraram um mesmo corte para todas as classes (sinalizado por P), segundo a constante de decaimento pequena, com diferentes cortes por classe (indicado por C), identificados através da Figura 5.41. É nítido o melhor desempenho de diferentes granularidades por classe, com eficiências superiores de $\approx 1,2$ (35 neurônios) a $\approx 2,0$ (30 neurônios) pontos percentuais.



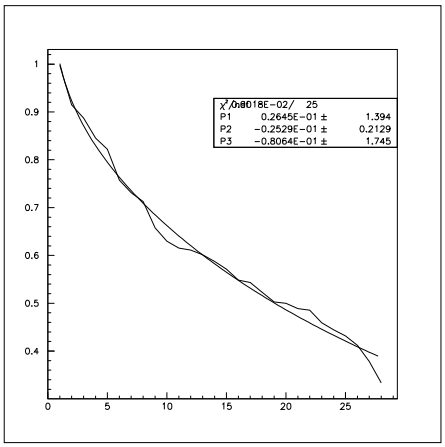
(a) Classe A



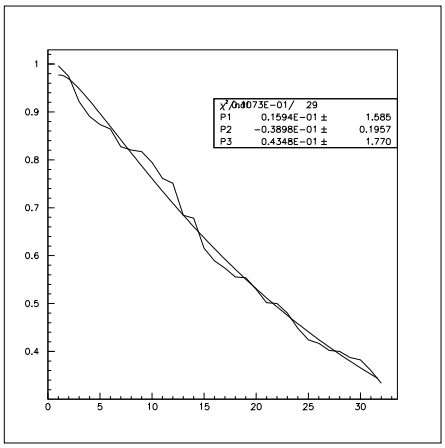
(b) Classe B



(c) Classe C

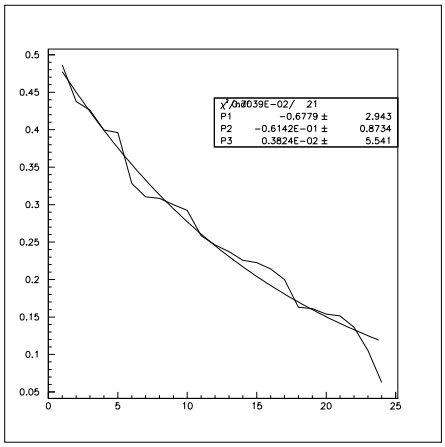


(d) Classe D

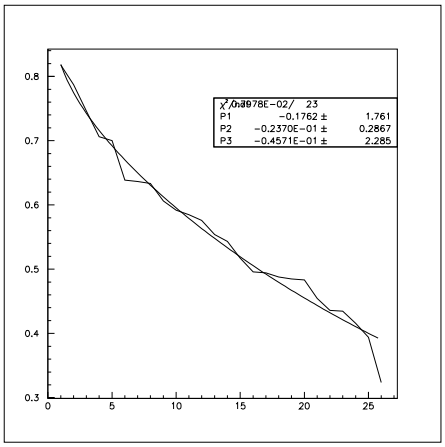


(e) Classe E1

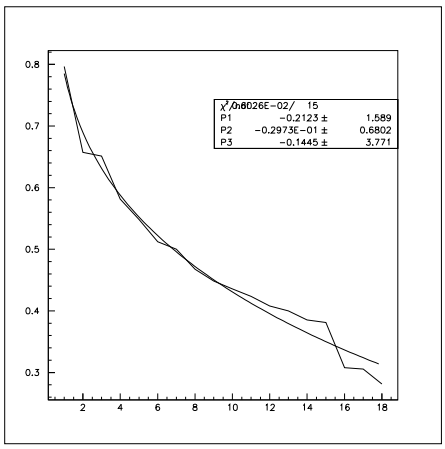
Figura 5.37: Modelagem da curva de dissimilaridade do agrupamento hierárquico, classe-a-classe, para vetores representativos definidos pelos vetores de pertinência das corridas aos grupos do agrupamento (classes A a E1).



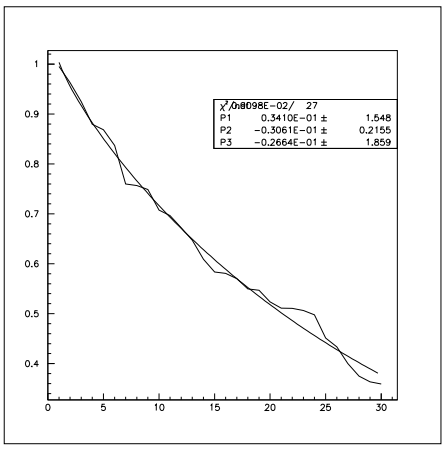
(a) Classe E2



(b) Classe F



(c) Classe G



(d) Classe H

Figura 5.38: Modelagem da curva de dissimilaridade do agrupamento hierárquico, classe-a-classe, para vetores representativos definidos pelos vetores de pertinência das corridas aos grupos do agrupamento (classes E2 a H).

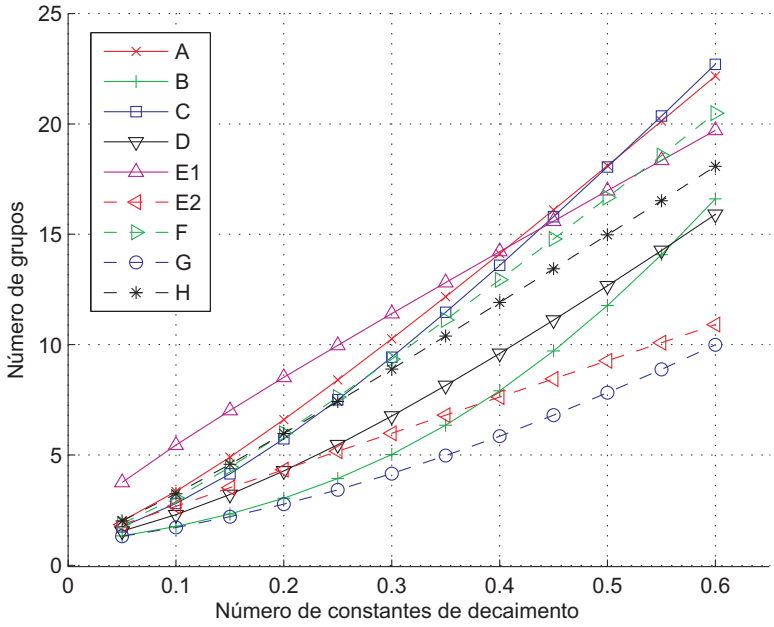
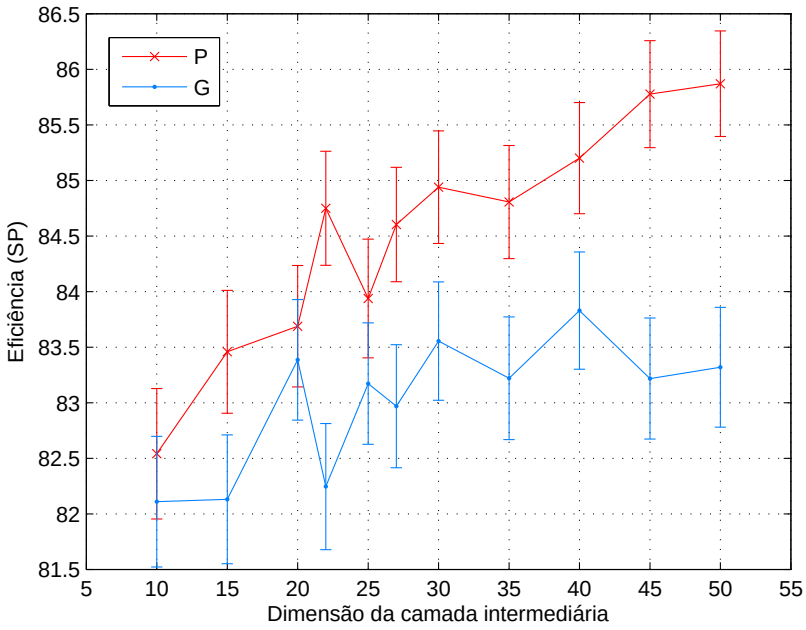
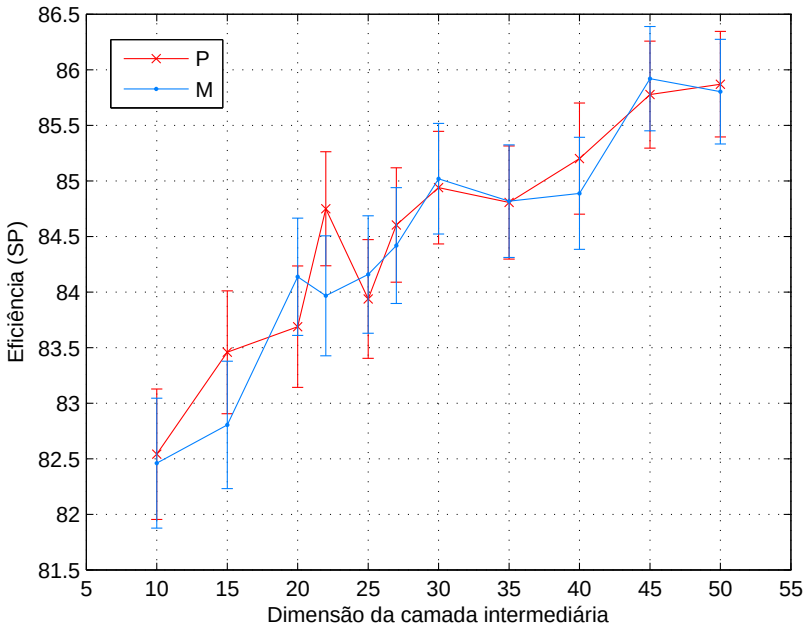


Figura 5.39: Número de grupos do dendrograma por classe e número de constantes de decaimento para o agrupamento hierárquico de vetores representativos definidos pelos vetores de pertinência das corridas aos grupos do agrupamento (Veja o texto).

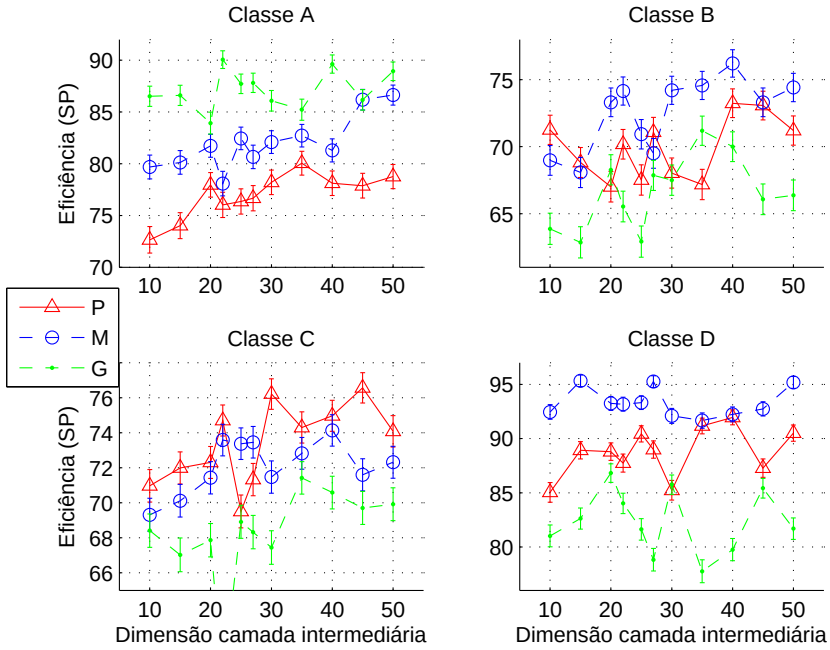


(a)

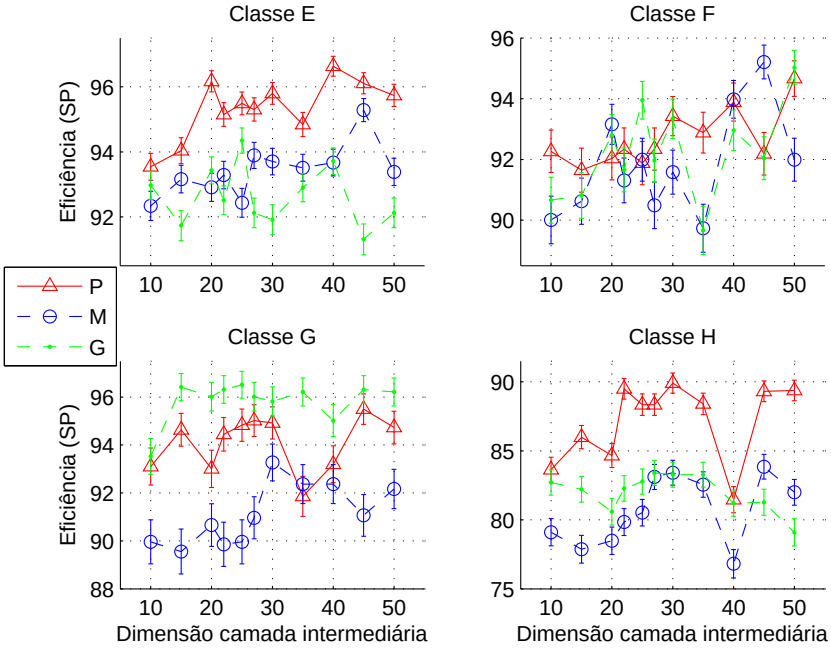


(b)

Figura 5.40: Eficiências de generalização (SP) dos classificadores produzidos com base no corte do agrupamento hierárquico de vetores representativos definidos pelos vetores de pertinência das corridas aos grupos do agrupamento, classe-a-classe, segundo mesmo nível de granularidade (Veja o texto).



(a)



(b)

Figura 5.41: Eficiências de generalização por classe e nível de granularidade, para uma seleção baseada em corridas, utilizando o agrupamento hierárquico, classe-a-classe, dos vetores de pertinência das corridas aos grupos do agrupamento (Veja texto).

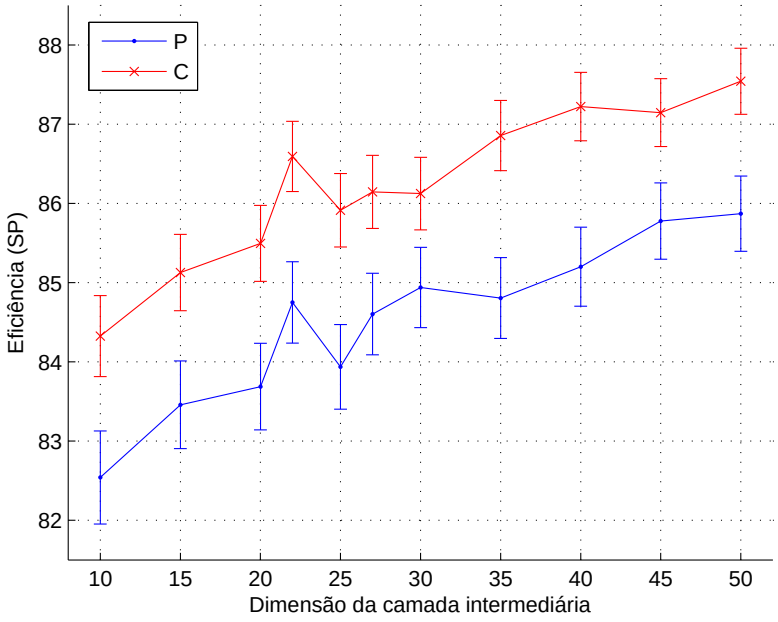


Figura 5.42: Eficiências de generalização dos classificadores obtidos através da seleção de corridas segundo uma mesma (P) ou diferentes granularidades para o corte das classes (C), considerando agrupamentos hierárquicos classe-a-classe dos vetores de pertinência das corridas aos grupos do agrupamento (Veja o texto).

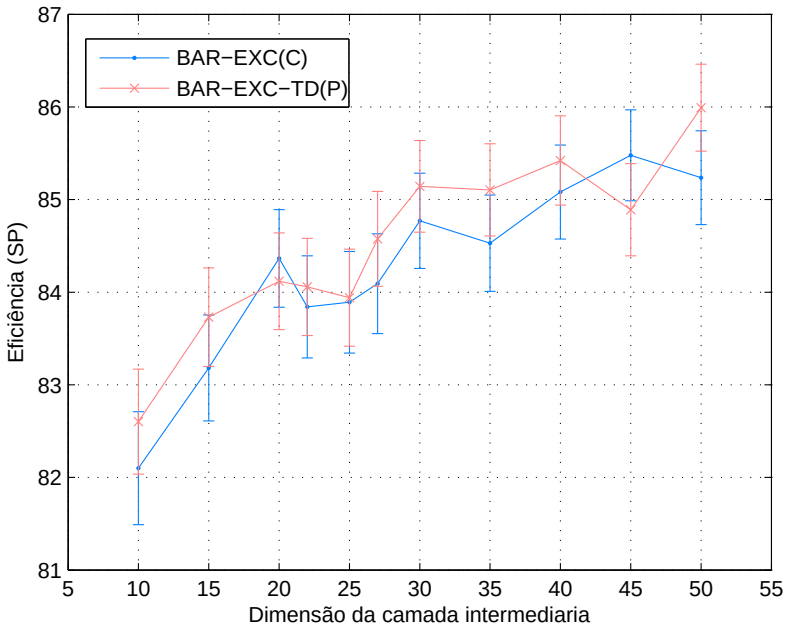
5.2.4.4 Comparação das diferentes propostas

Esta análise buscou identificar, dentre os critérios propostos, qual apresenta melhor desempenho, assim como comparar as propostas de seleção baseadas em agrupamentos com a baseada em subamostragem aleatória, de forma similar à análise realizada para a seleção baseada em espectros (vide seção 5.1.5.4).

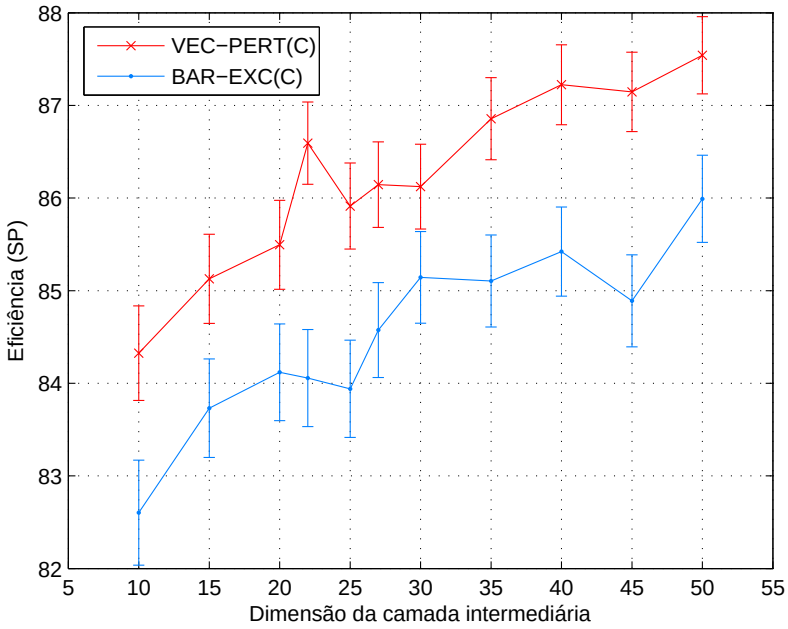
Inicialmente foram comparadas as seguintes modalidades: o agrupamento, classe-a-classe, dos baricentros dos centros excitados por cada corrida (BAR-EXC), utilizando diferentes granularidades por classe (C); o agrupamento dos baricentros dos centros excitados para todas as classe (BAR-EXC-TD), segundo a granularidade pequena; e, por fim, o agrupamento dos vetores de pertinência (VEC-PERT) com granularidades distintas por classe (C). A curva de eficiência SP associada a cada proposta é apresentada na Figura 5.43.

Comparadas as modalidades BAR-EXC e BAR-EXC-TD, o desempenho da primeira é ligeiramente superior, entre ≈ 0 (25 neurônios) a $\approx 0,6$ (35 neurônios) pontos percentuais, exceto para redes com 20 e 45 neurônios, o que sinaliza em favor de agrupamentos individuais, classe-a-classe, dos vetores representativos das corridas. Para as modalidades VEC-PERT e BAR-EXC, o desempenho da primeira é claramente superior, de $\approx 1,0$ (30 neurônios) a $\approx 2,6$ (22 neurônios) pontos percentuais. Deste resultado constata-se que a similaridade entre a excitação dos agrupamentos provê uma informação melhor para a seleção das corridas que os baricentros.

Passo seguinte consistiu em comparar a seleção VEC-PERT com a baseada em subamostragem aleatória, aqui referida como sorteio. A seleção VEC-PERT produziu classificadores de $\approx 1,3$ a $\approx 3,3$ pontos percentuais mais eficientes (SP) que a baseada em sorteio, a um custo computacional significativamente inferior.



(a)



(b)

Figura 5.43: Comparação das eficiências de generalização dos classificadores produzidos pela seleção de corridas através do agrupamento dos baricentros excitados classe-a-classe, com diferentes granularidades por classe (BAR-EXC(C)), dos baricentros excitados de todas as classes (BAR-EXC-TD(P)), com uma mesma granularidade por classe (P), e pelos vetores de pertinência (VEC-PERT(C)), com diferentes granularidades por classe (C) (Veja o texto).

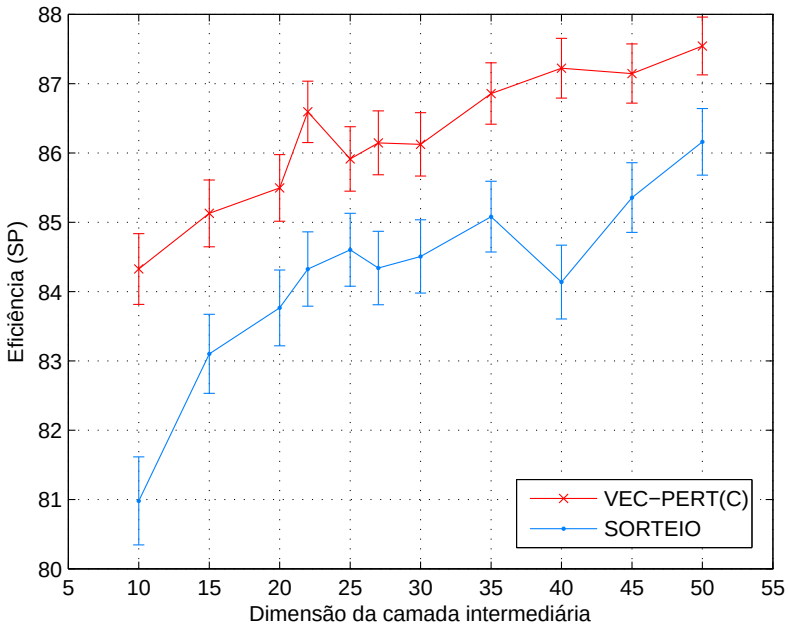


Figura 5.44: Comparação das eficiências de generalização dos classificadores produzidos pela seleção de corridas através da técnica proposta (VEC-PERT(C)) com a baseada em sorteio (SORTEIO).

Capítulo 6

Classificação baseada em múltiplos classificadores

Conforme discussões anteriores, num cenário real de operação, o sistema de classificação automática deverá lidar com um número significativamente maior de classes que as contempladas neste trabalho. No contexto do sonar passivo, propor um sistema classificador para a operação em cenários reais deve considerar, portanto, a utilização de técnicas que viabilizem a produção de um sistema capaz de discriminar, de forma eficiente, um número expressivo de classes, mesmo que desenvolvido sobre uma base de dados com restrições quanto à caracterização das classes. As limitações práticas existentes na aquisição dos sinais envolvidos e a multiplicidade de cenários existentes são os principais fatores responsáveis por estas restrições. Assim, é de especial importância que o sistema possua mecanismos para a detecção de eventos desconhecidos, informando ao operador quanto à sua ocorrência, e disponha da capacidade de adaptação a novos cenários, os quais são bastante prováveis.

Um enfoque atrativo para a constituição deste sistema é o dividir-e-conquistar, pelo qual o complexo problema da classificação é subdividido em problemas menores e mais simples. Uma estratégia interessante é realizar a integração de múltiplos classificadores, que podem explorar diferentes metodologias, escolhidas de forma a melhor atender aos requisitos e às especificidades das classes envolvidas. A combinação de vários classificadores resulta numa maior redundância com respeito às decisões, o que é útil para a segurança do sistema e, normalmente, se traduz em menores incertezas de classificação e melhor desempenho.

Em relação à identificação de cenários desconhecidos, este trabalho concentrar-se-á na identificação de novas classes. Uma vez identificada e apropriadamente caracterizada, é fortemente desejável que o sistema possibilite sua incorporação, de forma segura e eficiente, isto é, que sua adaptação ao incluir a nova classe comprometa, minimamente, o reconhecimento das classes já conhecidas.

O propósito deste capítulo é discutir a constituição do sistema classificador de contatos através de classificadores classe-escaláveis, metodologia que integra diferentes módulos, cada qual especializado na identificação de uma classe. Ênfase especial é dedicada às redes modulares classe-especialistas, para a qual é avaliada a identificação de eventos de classes desconhecidas e a incorporação de novas classes.

A estruturação deste capítulo é a seguinte: inicialmente, são discutidos múltiplos classificadores e suas abordagens mais comuns na literatura, que são as máquinas de comitê e os classificadores modulares. Em seguida, são apresentadas algumas técnicas de classificação que permitem a constituição de classificadores com uma fácil inserção de novas classes, aqui referidos como classe-escaláveis. Posteriormente, são apresentadas as redes modulares classe-especialistas, sendo discutidos critérios para o treinamento, definição da topologia e integração dos classificadores. Por fim, é discutida a detecção e inclusão de novas classes ao sistema proposto.

6.1 Múltiplos Classificadores

Para a solução de problemas complexos, que podem contemplar número expressivo de classes e/ou dados com características heterogêneas, a adoção de um único classificador pode originar soluções insatisfatórias [185], em especial, com respeito à capacidade de generalização [186]. Uma alternativa é realizar a combinação de múltiplos classificadores através de uma função apropriada, a qual deve enfatizar as potencialidades dos classificadores individuais e evitar suas fraquezas [185]. A constituição de um classificador através de diferentes redes neurais, por exemplo, pode resultar num desempenho superior ao modelo baseado em rede única [187], em especial, quando as características extraídas e a atuação das diferentes redes se complementam [185].

A idéia da combinação de classificadores remonta à antiga combinação de

estimadores, proposta por Laplace em 1818 [187]. O número de referências sobre este tema é abundante, havendo diferentes estratégias, interpretações e implementações. Não há, no entanto, uma teoria fechada sobre o assunto. A combinação de classificadores modifica o paradigma dos projetistas, antes preocupados em projetar um bom classificador com base em dados de dimensão reduzida, obtidos através de técnicas eficientes de compactação e extração de características. Como alternativa, podem ser produzidos classificadores baseados em vetores de características distintas, que atuem de forma complementar. Neste caso, ainda que cada classificador não possua uma eficiência excepcional, é possível obter um sistema de classificação de alta-performance [188].

Um princípio comumente explorado para múltiplos classificadores é o dividir-e-conquistar, que consiste em combinar vários classificadores para a solução de uma mesma tarefa, ou subdividi-la em tarefas mais simples, cada uma responsabilidade de um módulo particular. Para o primeiro caso, tem-se as máquinas de comitê [4], e para o segundo, os classificadores modulares [189]. A estratégia das máquinas de comitê é utilizar vários classificadores treinados para uma mesma tarefa, cujas saídas são combinadas visando produzir um sistema de maior eficiência que aquela obtida por um classificador individual. Os classificadores modulares, por outro lado, dividem sua tarefa em várias subtarefas mais simples, cada qual responsabilidade de um sistema ou modelo local especializado. É pressuposto, portanto, que exista uma divisão funcional ou estrutural considerável entre os diferentes módulos do sistema [189]. A seguir, será realizada uma revisão bibliográfica e discussão dos principais representantes dos dois enfoques.

6.1.1 Máquinas de Comitê

Em [190], Hansen e Salamon (1990) mostraram que numa mistura de N classificadores neurais, se todas as redes produzirem classificações corretas com verossimilhança de $(1-p)$, onde $0 \leq p \leq 1$, e cometerem erros independentes, para $p < \frac{1}{2}$ e um critério de decisão por voto majoritário, o erro de classificação é uma função monotonicamente decrescente com N, isto é, quanto maior o número de classificadores, menor será o número de erros cometidos.

Em problemas reais, a condição de independência dos erros dos diferentes

classificadores será dificilmente satisfeita. Diferentes estratégias são utilizadas para reduzir o nível de dependência dos classificadores, entre elas: o treinamento segundo diferentes parâmetros iniciais, a utilização de classificadores distintos ou seguindo uma mesma metodologia, porém com diferente número de parâmetros, ou ainda, a adoção de diferentes representações dos dados e conjuntos para o treinamento. Para algumas metodologias, mais de uma destas estratégias é utilizada simultaneamente.

Na linha de diferentes representações dos dados, Benediktsson et al. (1997) propõem as redes neurais paralelas consensuais [191]. Para esta arquitetura, cada classificador utiliza dados pré-processados pela transformada *wavelet* [24], segundo diferentes níveis de resolução. Uma alternativa, útil para vetores de dados de elevada dimensionalidade, deve-se a Schmidt (1996) [192], onde cada classificador considera apenas parte da dimensão dos vetores de dados. Segundo a estratégia de diferentes conjuntos de treinamento, a técnica de *Bagging*, proposta por Breiman (1996) [193], projeta classificadores com base em amostras *Bootstrap* [148]. Outras alternativas nesta linha são os algoritmos de *Boosting* [194] e *Adaboost* [195]. Uma proposta que combina várias das estratégias citadas anteriormente é o *Redundant Classification Environment (RCE)*, proposto por Shimshoni e Intrator (1998) [196]. Para o RCE, tem-se uma arquitetura de classificadores em níveis, onde cada nível explora uma estratégia diferente: no primeiro, são considerados classificadores idênticos treinados segundo diferentes valores de pesos iniciais. Para o segundo, são consideradas diferentes amostras *Bootstrap*. O último nível utiliza diferentes representações dos dados ou arquiteturas neurais. Em todas as propostas anteriores, a topologia dos classificadores deve ser definida a priori pelo projetista. Uma alternativa é o método de agrupamento construtivo de redes neurais (*Constructive NN Ensamble - CNNE*), proposto por Islan et al. (2003) [197], o qual propõe a constituição de sistemas classificadores baseados em comitês de redes MLP, cujo número e complexidade é definido através de um critério automático.

Tão relavante quanto definir, apropriadamente, os diferentes classificadores a serem utilizados no comitê, é prover um critério de decisão que considere, de forma eficiente, o conhecimento adquirido por cada um. Entre estratégias aplicáveis às máquinas de comitê, têm-se os critérios de votação, ponderados e bayesianos.

Entre os critérios de votação, a votação por maioria [198] está entre os mais

simples e de aplicação mais geral. A votação pode ainda explorar critérios mais conservativos, onde busca-se o consenso entre todos classificadores, ou mais relaxados, onde as classes são indentificadas com base numa concondância apenas parcial. Outra possibilidade é considerar os resultados obtidos para outras classes, além da vencedora, no processo de decisão. Neste caso, limiares podem ser estabelecidos, decidindo-se por uma classe apenas quando a diferença do número de votos entre a primeira e a segunda candidatas, por exemplo, forem superiores a um dado valor. Em caso contrário, classifica-se o evento como pertencente a uma classe desconhecida [199]. Outro critério, que pode ser interpretado como uma votação baseada na ordenação das respostas dos classificadores para as diferentes classes, é a contagem por borda [185]. A votação pode ainda ser complementada por outros critérios, como na proposta de Woods et al. (1997) [200], que, para os casos onde não há unanimidade entre os classificadores, identifica a classe com base no classificador de melhor desempenho, o qual é definido, para cada evento, por um critério específico.

Para os critérios baseados em votação, as decisões são tomadas considerando um mesmo nível de significância para os classificadores, o que não leva em consideração as diferenças de desempenho entre eles. Uma alternativa são os critérios ponderados, onde a cada classificador é atribuído um peso diferente na formação da decisão conjunta. Um dos primeiros trabalhos desta linha deve-se a Perrone e Copper (1993) [201], que abordou a aplicação de múltiplas redes neurais à problemas de regressão, realizada através da combinação linear de suas saídas através de pesos específicos. Entre trabalhos expressivos nesta linha, tem-se Hashem (1995) [187], Breiman (1996) [202] e Ueda (2000) [203].

Pela teoria bayesiana, propor uma regra de decisão conjunta envolvendo múltiplos classificadores demanda o cálculo das probabilidades das várias hipóteses, o qual deve considerar as informações fornecidas por cada classificador. Contribuição interessante de Kittler et al.(1998) [198] é mostrar que, caso as saídas dos classificadores sejam probabilísticas, as regras do máximo, do valor médio das saídas dos classificadores e da votação por maioria, que são intuitivas e extensivamente utilizadas, podem ser derivadas com base na teoria de decisão bayesiana.

6.1.2 Classificadores modulares

Conforme mencionado anteriormente, nas redes modulares há um conjunto de modelos locais especializados que são combinados. Cada modelo realiza uma função explícita, interpretável e relevante à solução do problema [189]. Há uma forte conexão entre as redes modulares e os sistemas biológicos. O sistema visual dos primatas, por exemplo, apresenta módulos especializados na deteção do movimento, forma e avaliação das cores, enquanto o sistema nervoso central processa as respostas deste módulos para o reconhecimento de objetos [204].

Entre as características relevantes das redes modulares ao problema de sonar passivo tem-se [189]:

- Escalabilidade

  No contexto de operação, é provável que o sistema de sonar passivo venha a se deparar com novos cenários de operação, que podem contemplar desde novas condições operativas de classes já conhecidas, até mesmo classes completamente novas. É necessário, portanto, que o sistema possua plasticidade [205]. Num classificador não-modular, a incorporação de conhecimento adicional exige o retreino de toda estrutura, o que é contra-indicado pela possibilidade da perda do conhecimento anteriormente adquirido, assim como pelo eventual custo computacional envolvido. Um forte atrativo das redes modulares é que a incorporação de conhecimento pode ser realizada através da adição de novos módulos e/ou pelo retreino de parte deles, o que é conhecido como aprendizado incremental [206].

- Aprendizado

  Aplicações complexas podem exigir diferentes tipos de conhecimento e técnicas de processamento. Sistemas modulares permitem grande flexibilidade quanto ao aprendizado, podendo reunir diferentes técnicas de classificação, tais como: sistemas neurais supervisionados e não-supervisionados, classificadores bayesianos, entre outros. É possível também o aprendizado de diferentes mapeamentos sem a ocorrência do esquecimento catastrófico[1] [207].

[1]Entende-se como esquecimento catastrófico a perda de parte expressiva do conhecimento envolvido na solução de uma tarefa quando o sistema é submetido ao aprendizado de uma outra

- Integração do conhecimento

  A modularidade facilita a integração do conhecimento a priori sobre o problema, resultando num melhor aprendizado e maior eficiência do sistema.

- Redução da complexidade dos modelos envolvidos

  Em redes modulares, tarefas complexas são subdivididas em tarefas mais simples, que demandam uma estrutura e processo de aprendizado menos complexos. Estruturas menos complexas são especialmente indicadas para dados sujeitos a restrições estatísticas, tais como os provenientes de sonar passivo. A solução de um problema complexo por vários subsistemas pode, no entanto, originar um sistema de complexidade global maior que a solução baseada em um sistema único.

Algumas propostas de classificadores modulares exploram a combinação de técnicas de aprendizado não-supervisionado e supervisionado. Em geral, o primeiro conjunto de técnicas é utilizado para a partição do espaço dos dados; enquanto o segundo, para a classificação dos eventos de cada partição. Bartfei (1994) [208] propõe a decomposição dos dados em diferentes grupos através da técnica ART [56, 205]. Para cada grupo, um classificador MLP é treinado, sendo a classificação final definida pelo classificador que apresentar o maior valor de saída, o que corresponde a maximizar a probabilidade a posteriori de detecção das classes [52]. Há outros trabalhos nesta linha [209, 210]. Em [211] e [212] são exploradas a decomposição dos dados em diferentes níveis de granularidade, realizada através de uma topologia formada por diferentes redes ART dispostas em cascata. Entre outros métodos baseados em partições do espaço dos dados, tem-se a mistura adaptativa de especialistas locais [213] e a mistura hierárquica de especialistas [214], que possuem forte conexão com os modelos gaussianos de mistura [4].

Em alguns trabalhos [215, 216] verifica-se a utilização de algoritmos evolutivos para a produção e integração das redes modulares. Neste caso, tanto as conexões internas de cada módulo, ou seja, sua estrutura, quanto suas interconexões com os demais módulos é otimizada pelo algoritmo evolutivo. O custo computacional destas propostas é, no entanto, alto, o que restringe sua aplicação a problemas de

tarefa.

dimensionalidade reduzida. Outra possibilidade para a integração dos classificadores modulares é o uso de lógica nebulosa, em especial, através de integrais nebulosas, conforme realizado em [217].

6.2 Classificadores classe-escaláveis

Prover classificadores que permitam uma fácil incorporação de novas classes é interessante ao problema de sonar passivo, visto que é provável o sistema deparar-se com novos cenários operativos. Metodologias de classificação onde a detecção é baseada na integração de um ou mais módulos associados a cada classe serão aqui referidos como classificadores classe-escaláveis. Presume-se, portanto, que esta categoria de classificadores, devido à especialização, permitiria uma fácil incorporação de novas classes, realizada através da inclusão de novos módulos, seguida por um eventual ajuste no critério de integração.

Serão discutidas a seguir três técnicas aplicáveis à constituição de classificadores classe-escaláveis: os filtros casados, a técnica de curvas principais e os classificadores neurais classe-modulares.

6.2.1 Filtros Casados

Uma técnica clássica e consagrada para a detecção de sinais é a filtragem casada [27]. Por esta técnica, é presuposto que tanto o sinal proveniente da classe quanto o ruído ambiente possuem uma distribuição estatística conhecida, tipicamente, gaussiana. A decisão é baseada no critério de máxima probabilidade a posteriori, isto é, de posse do sinal recebido, decide-se em favor da classe mais provável que tenha produzido o sinal. Filtros casados, por serem baseados na modelagem estatística das classes, podem facilmente incorporar novas classes.

Na formulação das equações de decisão, para a detecção de sinais digitais de k classes, são consideradas as seguintes hipóteses (H_j):

$$H_j : \qquad \mathbf{y} = \mathbf{s}_j + \mathbf{n}, \qquad 1 \leq j \leq k, \tag{6.1}$$

onde $\mathbf{y}$ representa o sinal recebido, o qual é formado pelo sinal $\mathbf{s}_j$, característico ao processo estocástico da j-ésima classe, e $\mathbf{n}$, que está relacionado ao processo do

ruído ambiente. Na abordagem usual deste técnica, é pressuposto que $\mathbf{s}_j$ possua uma distribuição gaussiana de média μ_j e covariância $\mathbf{\Sigma}_j$, assim como o ruído é tido como gaussiano e branco, de média nula e covariância $\frac{N_o}{2}\mathbf{I}$, onde $\mathbf{I}$ é a matriz identidade.

Para o problema do sonar, há apenas aquisições de $\mathbf{y}$ para cada hipótese, e nenhum conhecimento sobre as distribuições estatísticas de $\mathbf{s}_j$ e $\mathbf{n}$ estão disponíveis. Para lidar com este problema, propõem-se uma caracterização alternativa para as hipóteses, realizada segundo [218]:

$$H_j : \qquad \mathbf{y} = \mathbf{s}_j, \qquad 1 \leq j \leq K, \tag{6.2}$$

onde $\mathbf{s}_j$ é suposto proveniente de uma distribuição gaussiana de média $\boldsymbol{\mu}_j$ e covariância $\mathbf{\Sigma}_j$, ou seja, o ruído ambiente passa a ser incorporado na caracterização de cada classe. Entre os atrativos deste enfoque, tem-se que nenhuma suposição sobre o ruído é realizada, assim como os parâmetros $\mathbf{\Sigma}_j$ e μ_j podem ser facilmente estimados com base nas aquisições de $\mathbf{y}$. No trabalho [218] foi avaliada a classificação de sinais de sonar passivo, utilizando um conjunto de dados similar ao deste trabalho, no entanto, mais restrito quanto ao número de classes (4, no total). Foi discutida, também, a otimização dos filtros quanto ao custo computacional, realizada através de componentes principais e minoritárias [67]. Resultados expressivos foram obtidos, porém o classificador mostrou-se bastante sensível à formação dos conjuntos utilizados para o seu projeto e avaliação.

6.2.2 Curvas Principais

As curvas principais foram propostas por Hastie e Stuezle (1989) [58], consistindo numa generalização não-linear da análise de componentes principais. Segundo a definição de Hastie e Stuezle (HS), uma curva principal é uma curva suave, isto é, infinitamente diferenciável, auto-consistente [2], e que não intercepta a si própria. Desta forma, as curvas principais "passam pelo meio" da nuvem de pontos provenientes da distribuição estatística dos dados, fornecendo um bom "resumo" unidimensional dos sinais.

A Figura 6.1 ilustra a curva principal para um conjunto de dados arbitrário.

[2] Auto-consistência significa que a curva é definida pelo valor médio dos pontos do espaço que nela possuem projeção.

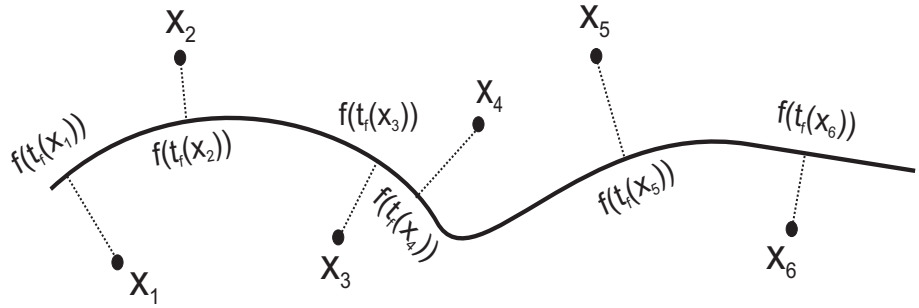


Figura 6.1: Modelo de curva principal

Seja $\mathbf{X} = (X_1, \ldots, X_d)$ um vetor aleatório de dimensão d, e $\mathbf{f}(t) = (f_1(t), \ldots, f_d(t))$ a curva principal em $\mathbf{R}^d$, com parâmetro $t \in \mathbf{R}^d$, definida de forma que, para todo $\mathbf{x} \in \mathbf{R}^d$, t assume um valor que minimiza a distância entre $\mathbf{x}$ e $\mathbf{f}(t)$.

Para reduzir a complexidade do processo de obtenção, as curvas são aproximadas, normalmente, por um número finito de segmentos de reta. Várias definições alternativas à proposta por HS são encontradas na literatura, entre elas: o algoritmo de linhas poligonais (PLA - *Poligonal Line Algorithm*), proposto por Kégl et at (2000) [219], o qual fixa o comprimento das curvas; um método que restringe a curvatura das curvas, proposto por Sandilya e Kulkarni (2002) [220]; e um método paramétrico, devido a Tibishirani (1992) [221], o qual define a curva principal com base na estimação do modelo de misturas, realizado através do algoritmo EM [49].

Dentre os métodos propostos, a técnica de k-segmentos não-suave, proposta por Verbeek et al. (2002) [59], se destaca em relação as demais, possuindo maior robustez na estimação das curvas, menor susceptibilidade a mínimos locais e convergência prática garantida. Segundo esta técnica, a curva principal é construída de forma incremental, ou seja, é iniciada com apenas um segmento, sendo este número aumentado progressivamente. A cada inserção de um novo segmento, toda curva é novamente otimizada.

As curvas principais são um poderoso sistema de extração de características não-lineares. Uma proposta é explorar esta técnica para a constituição de sistemas de classificação classe-escaláveis. Segundo esta proposta, produz-se uma curva característica à cada classe, e a classificação pode ser baseada em diferentes critérios, entre eles: índices de projeção ou de distância de eventos de entrada do classificador

às curvas. Em [222], um simples critério de decisão como a distância euclidiana dos eventos às curvas de cada classe mostrou-se bem sucedido na classificação de dados também provenientes de sonar passivo, considerando porém um número mais restrito de classes, navios e corridas.

No trabalho [223], a classificação por distância às curvas foi avaliada para o conjunto de dados utilizado neste trabalho. Discutiu-se, também, o impacto da complexidade da curva e da metodologia de normalização dos dados na eficiência de generalização do classificador. Os resultados obtidos mostraram que o classificador proposto possui baixo custo computacional em fase de operação, e seu desempenho, em termos de eficiência de detecção, foi bastante expressivo. A despeito de sua eficácia, o custo computacional da extração das curvas é elevado, e o algoritmo de extração é complexo, o que compromete a implementação de classificadores baseados em curvas que possuam a capacidade de adaptação em operação. Assim, este trabalho enfatizou o desenvolvimento de redes neurais classe-modulares, mais adequadas à constituição destes classificadores.

6.3 Classificadores neurais classe-modulares

Classificadores modulares formados por um conjunto de redes neurais especializadas na detecção de cada classe serão, aqui denominados, como classe-modulares. Para um classificador de K classes, uma possibilidade é utilizar K subclassificadores, cada um constituído por uma rede MLP [4] de duas camadas, com único neurônio na camada de saída. Cada subclassificador é especializado na detecção de uma classe, ou seja, é treinado para identificar se o evento pertence ou não à classe a ele associada [224]. Neste caso, de posse das saídas dos diferentes especialistas, cabe a uma unidade de decisão o papel de definir à qual classe um evento pertence. Esta arquitetura será aqui referida como sistema classificador neural classe-especialista, sendo ilustrada na Figura 6.2. Algumas aplicações desta proposta são [225], no reconhecimento de sinais de descargas parciais, e em [226, 227], na identificação de objetos desconhecidos em seqüências de vídeo.

Outra possibilidade para a constituição de sistemas classificadores classe-especialistas é a arquitetura MIN-MAX [228]. Nesta proposta, a classificação de K

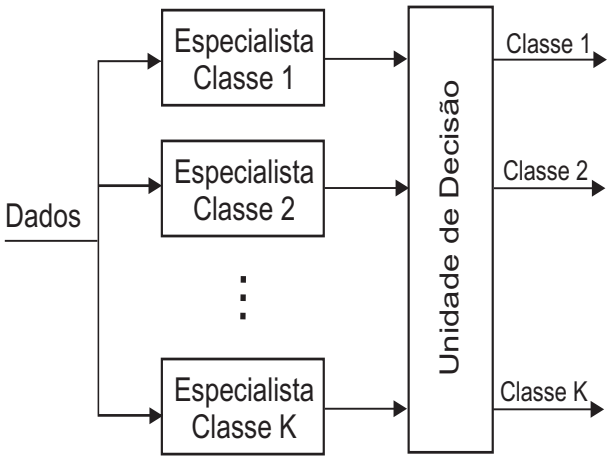


Figura 6.2: Sistema classificador neural classe-especialista

classes é subdividida em $K(K-1)$ classificadores que envolvem 2 classes arbitrárias i e j, para $1 \leq i \neq j \leq K$. Num problema de 3 classes, por exemplo, 6 módulos seriam produzidos, os quais compreenderiam a separação dos pares de classes $\{C_1, C_2\}$, $\{C_1, C_3\}$, $\{C_2, C_1\}$, $\{C_2, C_3\}$, $\{C_3, C_1\}$ e $\{C_3, C_2\}$, ainda que o módulo $\{C_3, C_2\}$, entre outros, possa ser constituído através da manipulação matemática da saída do módulo responsável pelo par $\{C_2, C_3\}$. No treinamento do classificador associado às classes $\{C_2, C_3\}$, por exemplo, os dados da classe 1 não seriam utilizados. Para tarefas de classificação complexas, é possível ainda explorar um maior nível de fragmentação, subdividindo cada módulo em subclasses. Nesta proposta, a integração dos módulos é realizada em dois níveis: no primeiro, K módulos formam um conjunto responsável pela identificação de cada classe, os quais são integrados pelo operador de mínimo; para o segundo é utilizado o operador de máximo, que define a classe através do conjunto de módulos que apresentar maior valor de saída. Nos trabalhos [229] e [230], esta arquitetura é utilizada para a classificação de sinais de eletroencefalograma e para o reconhecimento de locutor através do discurso.

Em razão do menor número de classificadores envolvidos, e por uma maior flexibilidade quanto aos critérios de integração, quando comparado com o critério MIN-MAX, este trabalho se concentrará nas redes classe-especialistas, sendo discutidos, a seguir, critérios para o treinamento, definição das topologias dos especialistas e sua integração.

6.3.1 Treinamento das redes Classe-Especialistas

No sistema de redes neurais classe-especialistas, módulos treinados de forma independente são integrados e atuam de forma cooperativa. Como, neste sistema classificador, os diferentes módulos irão competir pelos eventos, é desejável que cada especialista possua um desempenho expressivo na detecção de sua classe, assim como a capacidade de rejeitar, de forma eficiente, os eventos pertencentes às demais classes.

Dois aspectos relevantes no treinamento de uma rede neural são a função objetivo e o critério de parada: o primeiro, relacionado diretamente com as características extraídas pelo classificador; e o segundo, que visa evitar o excesso de treinamento e a conseqüente perda da capacidade de generalização. A escolha destes parâmetros, considerando as especificidades das redes classe-especialistas, será discutida a seguir.

6.3.1.1 Escolha da função objetivo

Usualmente, o treinamento de classificadores MLP utiliza o erro médio quadrático como função objetivo. Para um problema com K classes, com n_i eventos por classe no conjunto de treino, esta função pode ser escrita como [3]:

$$E = \frac{1}{n_t} \sum_{j=1}^{K} \sum_{i=1}^{n_j} (\mathbf{t}_i^j - \mathbf{z}_i^j)^2, \tag{6.3}$$

para:

$$n_t = \sum_{l=1}^{K} n_l, \tag{6.4}$$

onde $\mathbf{z}_i^j$ e $\mathbf{t}_i^j$ correspondem aos i-ésimos vetores de saída e alvo da j-ésima classe, e n_t é o número total de eventos disponíveis. Esta função determina um erro médio para todos eventos e classes, constituído, em sua maior parte, pelas classes mais populosas, as quais tendem a ser "beneficiadas"no treinamento.

Num sistema classificador classe-especialista, um classificador binário é dedicado a cada classe, havendo, portanto, K classificadores. Neste caso, a Equação 6.3,

[3]Como o algoritmo RPROP [110] foi utilizado para o treinamento das redes classe-especialistas, serão apenas consideradas funções objetivo que presumem um treinamento na modalidade batelada [4] para os classificadores.

para o p-ésimo classificador ($1 \leq p \leq K$), pode ser escrita como:

$$\begin{aligned} E_p &= \frac{1}{n_t}\sum_{i=1}^{n_i}(t_i^p - z_i^p)^2 + \frac{1}{n_t}\sum_{j\neq p, p=1}^{K}\sum_{i=1}^{n_i}(t_i^j - z_i^j)^2, \qquad 1 \leq p \leq k, \\ E_p &= \alpha_c E_c^p + \alpha_{nc} E_{nc}^p, \end{aligned} \tag{6.5}$$

para:

$$\alpha_c = \alpha_{nc} = \frac{1}{n_t}, \tag{6.6}$$

onde são evidenciadas duas parcelas na formação do erro: a primeira (E_c^p), que está associada à detecção da própria classe, aqui referida como erro da classe; e a segunda (E_{nc}^p), relacionada à rejeição das demais classes, aqui referida como erro da não-classe. As constantes α_c e α_{nc} respondem pelos pesos atribuídos aos erros da classe e da não-classe na constituição de E_p.

Em relação aos vetores alvos, suposição natural é $t_i^p = 1$ e $t_i^j = -1$ ($j \neq p$), ou seja, o módulo especialista deve apresentar saída $+1$ para os eventos da classe a ele associada; e -1, para as demais classes. Erros cometidos na parcela E_c^p respondem por eventos pertencentes à classe do especialista por ele não identificados, logo contribuem para a probabilidade de perda do alvo (contato) [27, 52]. Erros em E_{nc}^p correspondem a eventos classificados como pertencentes à classe do especialista, porém oriundos de outras classes, os quais contribuem para a probabilidade de falso-alarme [27, 52]. Aspecto desejável no treinamento do especialista é maximizar a probabilidade de detecção e minimizar o falso-alarme, o que corresponde a minimizar, conjuntamente, E_c^p e E_{nc}^p. Esta minimização pode ser realizada através da Equação 6.5, a qual considerou um mesmo peso para ambas as parcelas.

Ainda que as classes possam possuir um número equivalente de eventos, ao realizar o treinamento pela Equação 6.5, num problema com K classes, o conjunto associado à não-classe possui um número de eventos $(K-1)$ vezes maior que o relativo à classe. Tratam-se, portanto, de conjuntos desbalanceados [231], onde a estatística disponível para a caracterização da não-classe tende a ser significativamente maior que a da classe. Este fato resulta num treinamento que tende a privilegiar a rejeição de eventos da não-classe em detrimento da detecção dos eventos da classe, o que pode resultar numa detecção insatisfatória das classes, em especial, para classificadores com valores de K expressivos, tais como o de sonar passivo. Um melhor

compromisso entre a caracterização da classe e da não-classe pode ser obtido pela alternativa:

$$E_p \quad = \alpha_c E_c^p + \alpha_{nc} E_{nc}^p, \tag{6.7}$$

para:

$$\alpha_c = \frac{1}{2}\left(\frac{1}{n_p}\right) \tag{6.8}$$

$$\alpha_{nc} = \frac{1}{2}\left(\frac{1}{\sum_{l\neq p, p=1}^{K} n_l}\right), \tag{6.9}$$

ou seja, determinar o valor de E_p através do valor médio dos erros associados à classe e à não-classe. Para ambas propostas, o gradiente da função objetivo pode ser escrito como:

$$\nabla E_p = \alpha_c \nabla E_c^p + \alpha_{nc} \nabla E_{nc}^p, \tag{6.10}$$

isto é, tendo como base uma combinação ponderada de um gradiente médio associado à classe (∇E_c^p) e à não-classe (∇E_{nc}^p). Para a proposta anterior (Equação 6.7), a combinação utiliza o valor médio dos gradientes, a qual pode, no entanto, ser dominada pelo gradiente de maior módulo.

Uma alternativa, inicialmente contemplando classificadores binários [231], e posteriormente aplicada às redes classe-especialistas [224], considerou a formação deste gradiente através de:

$$\nabla E_p = \frac{1}{2}\left(\frac{\nabla E_c^p}{||\nabla E_c^p||} + \frac{\nabla E_{nc}}{||\nabla E_{nc}^p||}\right), \tag{6.11}$$

que é independente dos módulos de cada gradiente (classe e não-classe), e consiste no vetor bissetor do plano por eles definido. A adoção deste gradiente corresponde à utilização da seguinte função objetivo:

$$E_p = \frac{1}{2}\left(\frac{1}{||\nabla E_c^p||}\right) E_c^p + \frac{1}{2}\left(\frac{1}{||\nabla E_{nc}^p||}\right) E_{nc}^p, \tag{6.12}$$

ou seja, em escolher α_c e α_{nc} para a Equação 6.7 dados por:

$$\alpha_c = \frac{1}{2}\left(\frac{1}{||\nabla E_c^p||}\right) \tag{6.13}$$

$$\alpha_{nc} = \frac{1}{2}\left(\frac{1}{||\nabla E_{nc}^p||}\right), \tag{6.14}$$

que define uma função objetivo que se altera a cada passo de treinamento, de forma a estabelecer diferentes compromissos entre a informação da classe e da não-classe.

As modalidades propostas pela Equações 6.5, 6.7 e 6.12 serão aqui referidas como batelada uniforme (BU), classe-e-não-classe (CNC) e classe-e-não-classe normalizada (CNC-N), respectivamente, e serão avaliadas, posteriormente, no treinamento dos especialistas.

6.3.1.2 Definição do critério de parada

Conforme discutido na seção 4.1, para evitar uma demasiada especialização do classificador neural no conjunto de treinamento, e conseqüente perda da capacidade de generalização, deve-se prover algum mecanismo implícito ou explícito para o controle do aprendizado. O critério utilizado neste trabalho é o da parada antecipada [4], que se baseia na evolução do valor da função objetivo para um conjunto de eventos independentes (conjunto de avaliação).

Para uma melhor compreensão, na Figura 6.3 é ilustrada a evolução do valor da função objetivo de um classificador arbitrário para dois conjuntos: o de projeto e o de avaliação. É possível observar que o erro associado ao conjunto de projeto é uma função monotonicamente descrescente com respeito ao número de passos. Em relação ao conjunto de avaliação, o erro é, inicialmente, decrescente até a iteração sinalizada pelo marcador B, a partir da qual o erro se torna crescente. Este ponto indica quando o classificador começa a perder sua capacidade de generalização, visto o aumento do erro do conjunto de avaliação, sinalizando o momento de parada do treinamento.

Da observação da figura anterior, é possível verificar que, na faixa de iterações entre A e B, para a curva associada ao conjunto de avaliação, o classificador apresenta, do ponto de vista do erro quadrático, um ganho de generalização pequeno, praticamente nulo. Para o conjunto de projeto, na mesma faixa, a evolução do erro também é lenta, tendendo a se estagnar, o que significa que o classificador pode ser considerado treinado nesta faixa.

A cada iteração do treinamento, um conjunto de parâmetros (pesos e limiares) é proposto para o classificador. Pela curva da Figura 6.3, do ponto de vista do erro quadrático, a faixa AB provê parâmetros que resultam em classificadores

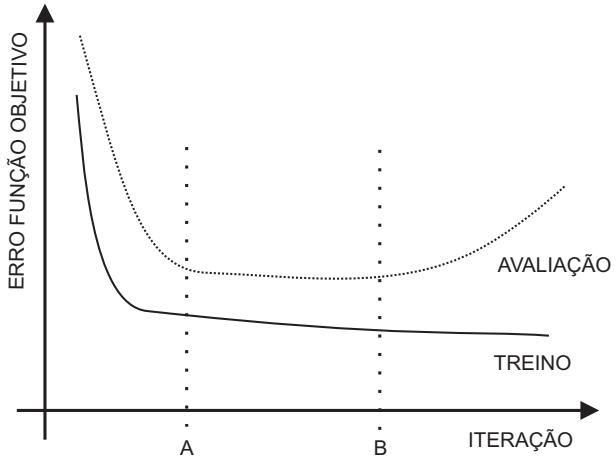


Figura 6.3: Evolução do erro associado ao conjunto de projeto e avaliação de um classificador neural arbitário

equivalentes. Observou-se, no entanto, para o treinamento de classificadores classe-especialistas de diferentes classes, variações expressivas da probabilidade de detecção e de falso-alarme [4] nesta faixa. Cabe observar que a detecção e o falso-alarme são índices de desempenho simples, porém bastante úteis à avaliação de classificadores classe-especialistas. Pelo primeiro, avalia-se a capacidade do classificador em reconhecer a classe a que está associado; pelo segundo, de rejeitar eventos de outras classes. Uma reduzida taxa de falso-alarme para esta modalidade de classificadores se traduz, normalmente, numa baixa confusão entre as classes no sistema final. Assim, para diferentes iterações na faixa AB foram verificados diferentes compromissos entre a capacidade de detecção e rejeição dos especialistas. Este fato motivou a utilização de diferentes índices de desempenho para prover a identificação do melhor passo de treinamento desta faixa.

Como critérios de escolha do melhor passo de treinamento, foram considerados alguns dos índices de desempenho discutidos na seção 4.2.1. Uma primeira avaliação realizou a seleção pelo maior valor da probabilidade de detecção da classe (PD) ou pela menor probabilidade de falso-alarme (PF). Em ambos casos, apenas

[4]Para a determinação das probabilidades de detecção e de falso-alarme foi considerado um limiar de decisão em zero, suposição razoável para viabilizar uma avaliação individual dos classificadores classe-especialistas, tendo em vista que, conforme citado anteriormente, são treinados para apresentar os valores -1 e +1 para eventos da não-classe e classe, respectivamente.

uma habilidade do classificador é avaliada: a de identificação ou rejeição de eventos, respectivamente. Para permitir uma avaliação onde ambas habilidades fossem consideradas, conjuntamente, foram também consideradas a área da ROC e a eficiência SP, a última determinada através de:

$$M_A = \frac{1}{2}(P_D + (1 - P_F)) \tag{6.15}$$

$$M_G = \sqrt{P_D(100 - P_F)} \tag{6.16}$$

$$SP = \sqrt{M_A M_G}, \tag{6.17}$$

para P_D e P_F expressos em percentual, buscando-se, portanto, um equilíbrio entre a maximização de P_D e a minimização de P_F. Por fim, foi avaliado realizar a parada através do valor mínimo da função objetivo, aqui referida como parada por erro quadrático (EMQ).

6.3.2 Escolha da topologia

Para a constituição dos módulos classe-especialistas, serão consideradas redes MLP de duas camadas. Faz-se, necessário, portando, definir o número de neurônios na camada intermediária de cada classificador, de forma que os classificadores possuam uma complexidade compatível com o conteúdo estatístico dos dados e o problema. Assim, a topologia não deve deve crescer além do necessário, em especial, no problema de sonar, tendo em vista as restrições estatísticas existentes no conjunto de dados.

Um conjunto de estratégias para a definição da topologia dos especialistas são as técnicas construtivas. Por estas técnicas, redes de complexidade crescente são construídas, sendo selecionada, através de um índice de desempenho específico, a rede que resultar numa maior capacidade de generalização (validação cruzada). Em outras palavras, busca-se, entre os modelos avaliados, o de complexidade mínima, que possua, no entanto, a melhor capacidade de generalização.

Algumas técnicas construtivas aplicáveis às redes MLP são a técnica de criação dinâmica de nós (*DNC - Dynamic Node Creation*) [232]; o treinamento PCD construtivo (PCD) [158], que é descrito no apêndice B, e a proposta de construção seqüencial de redes (SNC) [233]. Basicamente, a diferença entre as técnicas é a ocorrência ou não do congelamento de parâmetros (pesos e limiares). Nas três pro-

postas, o processo é iniciado com uma rede de um único neurônio. Finalizado seu treinamento, inicia-se o segundo estágio, onde mais um neurônio é inserido, e o treinamento é reiniciado. Este processo é repetido por mais estágios, até a rede possuir um número arbitrário de neurônios. Pelas propostas PCD e SNC, diferentemente da DNC, na inserção de um novo neurônio, os pesos associados a outros neurônios da camada intermediária não são alterados. Concluído o treinamento num estágio, a técnica SNC realiza um reajuste de todos os pesos; enquanto a PCD, não o faz. Em [2], as técnicas DNC e PCD mostraram resultados similares para a classificação de sinais de sonar.

Este trabalho considerará a técnica PCD construtiva para a definição da topologia dos especialistas. De forma similar à análise realizada para a seleção do melhor passo de treinamento, esta escolha considerará, também, o valor de 5 índices de desempenho (PF, PD, ROC, SP e EMQ) estimados com base no conjunto de avaliação, e será discutida, em maiores detalhes, na seção 6.4.2.

6.3.3 Integração dos classificadores

Tão relevante quanto dispor de módulos eficientes na identificação das classes, o que demanda uma apropriada seleção da função objetivo e do melhor passo de treinamento, é prover um critério que realize uma criteriosa integração das decisões individuais para a constituição da decisão final.

Um critério comumente utilizado [224, 225, 226, 227] é o máximo, que consiste em definir a classe a qual o evento pertence através do especialista de maior valor de saída. Conforme discussão realizada em [198], este critério corresponde à maximizar a probabilidade a posteriori de detecção das classes. Um dos atrativos do critério de máximo é a sua simplicidade, não demandando o ajuste de parâmetros para a constituição do sistema integrador, ou ainda, seu reajuste, caso novas classes sejam inseridas ao sistema.

Pelo critério do máximo, as decisões individuais são tratadas segundo um mesmo peso na constituição da decisão final, ainda que os classificadores possam apresentar diferenças de desempenho. Uma alternativa é considerar, através de combinadores lineares, diferentes pesos para as saídas de cada especialista. Seja um problema com K classes e um conjunto de dados com N eventos. Uma primeira

possibilidade, aqui referida como critério linear um, é produzir, com base na saída do j-ésimo especialista y_j^i $(1 \leq j \leq K)$, supostamente alimentado pelo i-ésimo evento do conjunto de dados $(1 \leq i \leq N)$, K variáveis, referidas como z_j^i, cada uma associada à detecção de uma classe, segundo a equação:

$$z_j^i = \sum_{l=1}^{K} \alpha_{jl} y_l^i, \tag{6.18}$$

onde α_{jl} são coeficientes ajustados de forma que z_j^i apresente o valor +1, caso o i-ésimo evento pertença à j-ésima classe; e -1, em caso contrário.

Para obter os coefientes α_j, pode-se formular o problema da seguinte forma:

$$\mathbf{M}_j \mathbf{w}_j = \mathbf{t}_j, \tag{6.19}$$

onde $\mathbf{M}_j$ é uma matriz, de dimensões N x K, onde cada linha corresponde às saídas dos especialistas para um dos N eventos de entrada possíveis, e $\mathbf{w}_j$ e $\mathbf{t}_j$ são vetores coluna, o primeiro, com K componentes, determinadas pelos coeficientes α_j da Equação 6.18; e o último, com N componentes, onde cada componente assume o valor +1, caso o evento a ela associada pertença à classe; ou -1, em caso contrário.

A equação 6.19 provê k sistemas lineares sobredeterminados, cuja solução de erro quadrático mínimo pode ser determinada, analiticamente, através de [67]:

$$\mathbf{w}_j = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{t}_j = \mathbf{M}^+ \mathbf{t}_j, \tag{6.20}$$

onde $\mathbf{M}^+$ é a pseudo-inversa da matriz M.

Um outro critério de integração, aqui referido como linear dois, consiste em determinar z_j^i pela fórmula:

$$z_j^i = \sum_{l=1}^{K} \alpha_{jl} y_l^i + \beta_j, \tag{6.21}$$

cuja principal diferença em relação à Equação 6.19 é dispor de um termo constante β_j, que também deve ser ajustado. Neste caso, uma forma simples para obter α_{jl} e β_j é realizar o treinamento de uma rede neural de uma única camada, com K neurônios lineares, cujas entradas são alimentadas pelas saídas dos especialistas, e os vetores alvos são definidos utilizando uma codificação maximamente esparsa das classes [4].

Outra alternativa consiste em combinar, de forma não-linear, as cararacterísticas extraídas por cada especialista. Neste caso alimenta-se a rede neural integradora com as saídas da camada intermediária dos especialistas de cada classe, eliminando-se suas camadas de saída. Assim, cada classificador especialista atua como um sistema de extração de características relevantes à discriminação da classe a ele associada, cabendo a rede integradora o papel de combinar as informações discrimantes providas por cada módulo para a tomada da decisão final. Este critério de integração será aqui referido como neural.

6.4 Resultados para a classificação baseada em redes classe-especialistas

Para avaliar a aplicação de classificadores classe-especialistas ao problema de sonar, a seguir serão apresentados os resultados referentes à escolha da função objetivo, do critério de seleção do melhor passo e do processo de integração dos especialistas. Nestas análises, buscou-se identificar qual conjunto de escolhas resulta numa maior eficiência de generalização para o sistema. Para o projeto e avaliação dos classificadores envolvidos, a fim de permitir um teste mais robusto e realístico para o sistema, foi considerada uma seleção por corridas, conforme critério da seção 5.2.2. Para o treinamento das redes neurais, de forma similar aos classificadores do Capítulo 5, foi utilizado o algoritmo RPROP [110].

6.4.1 Escolha da função objetivo e do critério de parada

Nesta análise foram consideradas as 3 funções objetivo descritas na seção 6.3.1.1: batelada uniforme (BU), classe-e-não-classe (CNC) e classe-e-não-classe normalizada (CNC-N), e 5 critérios de parada: detecção (PD), falso-alarme (PF), eficiência SP (SP), área da ROC (ROC) e erro quadrático médio (EQM), discutidos na seção 6.3.1.2, para o treinamento das redes envolvidas no sistema classe-especialista. Foram produzidos um total de $3.5 = 15$ sistemas classificadores, cada um baseado em 8 redes classe-especialistas, treinadas segundo uma mesma função objetivo e um mesmo critério de seleção do melhor passo. O critério do máximo foi utilizado para a integração das redes. Em razão do problema de mínimos locais [4], para a definição

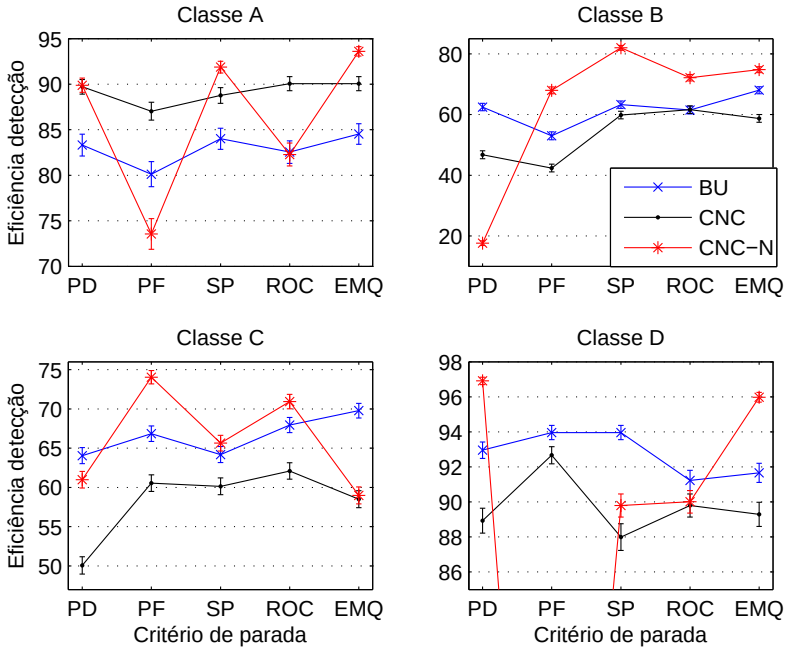
de cada rede, foi realizado o treinamento de 10 redes neurais, segundo diferentes inicializações, sendo escolhida a rede de melhor desempenho. Para esta seleção foi utilizado o mesmo índice considerado na escolha da iteração de parada.

Tendo em vista que a definição do número de neurônios da camada intermediária é dependente da escolha da função objetivo e do critério de parada, os resultados apresentados correspondem a especialistas com complexidade mínima, isto é, com 557 nós de entrada, apenas um neurônio na camada intermediária e 1 neurônio de saída, para os quais é esperado um comportamento similar ao que seria verificado no sistema final, onde cada especialista possuirá um número arbitrário de neurônios, que será definido posteriormente.

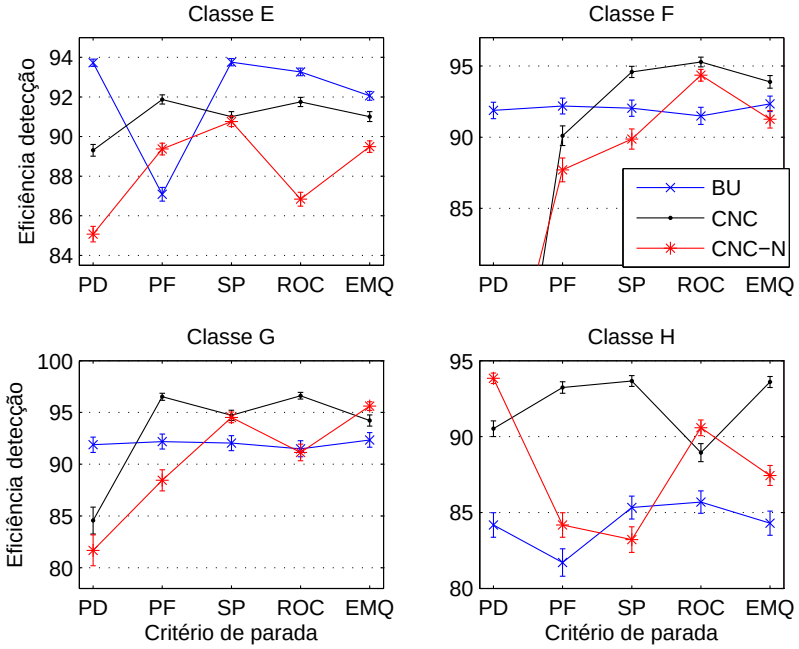
A faixa de iterações para a análise foi determinada da seguinte maneira: define-se a última iteração da faixa (ponto B da Figura 6.3) pelo mínimo da função objetivo para o conjunto de avaliação. Para esta iteração, identifica-se o valor correspondente da mesma função objetivo para o conjunto de projeto, o qual baseia a escolha da primeira iteração, definida de forma a corresponder a um valor superior em 10% deste valor.

As eficiências de detecção, obtidas para cada classe, considerando sistemas classificadores produzidos segundo diferentes funções objetivo e critérios de parada são apresentados na Figura 6.4. Pode-se verificar que, para a maior parte das classes, o gradiente CNC-N é o mais sensível ao critério de seleção do melhor passo; enquanto o BU, em geral, é o menos sensível. Para as classes A, B, C, D e H, uma maior eficiência de detecção é obtida pela função CNC-N para as seleções por PD (classes D e H), PF (classe C), SP (classe B) e EMQ (classe A). Para as classes F e G, tem-se a função CNC com seleção por ROC; e para classe E, a função BU com seleção por PD. Para o gradiente CNC-N, as seleções por PD e PF apresentam, em geral, um desempenho ruim quando comparadas às demais opções, como, por exemplo, para as classes B, E, F e G (PD) e A, D e H (PF).

Da análise anterior, verificou-se um melhor desempenho geral da função CNC-N, porém não ficou evidente qual seria o melhor critério de parada. Para auxiliar esta definição, foram avaliadas a eficiência média, apresentada na Figura 6.5(a); e a eficiência SP, na Figura 6.5(b), dos sistemas classificadores produzidos segundo diferentes escolhas da função objetivo e do critério de parada. Através dos



(a)



(b)

Figura 6.4: Eficiências de detecção, classe-classe, por critério de parada e função objetivo explorados no treinamento das redes classe-especialistas (veja o texto).

dois índices (média e SP) foi possível verificar um melhor desempenho dos classificadores treinados através dos gradientes CNC-N, com parada realizada pelos índices SP, EMQ e ROC, respectivamente, quanto ao desempenho global do sistema. Em primeiro lugar, tem-se o critério por SP, com eficiência média de $(86,0 \pm 0,3)\%$, e SP de $(85,7 \pm 0,5)\%$; em segundo, o EMQ, com eficiências de $(84,8 \pm 0,3)\%$ (média) e $(85,4 \pm 0,5)\%$ (SP); e por fim, o ROC, com eficiências de $(85,9 \pm 0,3)\%$ (média) e $(84,6 \pm 0,5)\%$ (SP). As eficiências classe-a-classe para estes índices são resumidas na Tabela 6.1. É possível perceber que, para todos os critérios, a classe C é a de detecção mais crítica. Para as seleções SP e EMQ, o desempenho é bastante similar, ligeiramente superior para a primeira, com destaque as classes B e E.

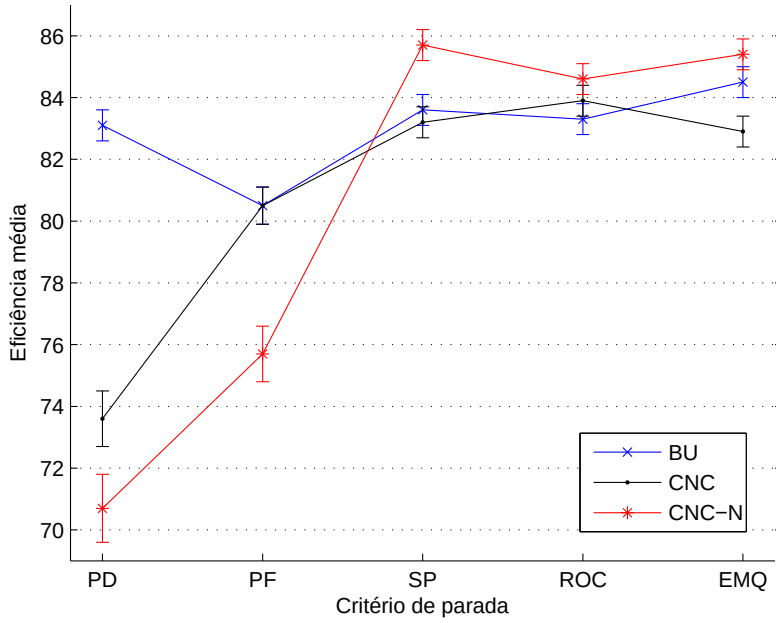
Tabela 6.1: Eficiências classe-a-classe (%), média (%) e SP (%) para o sistema especialista com treinamento por função objetivo CNC-N para diferentes critérios de parada.

Critério	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Eficiência média (%)**	**Eficiência SP (%)**
SP	91,9	82,0	65,6	89,8	90,8	89,9	94,5	83,2	$86,0 \pm 0,3$	$85,7 \pm 0,5$
ROC	82,3	72,1	70,9	90,0	86,8	94,3	91,1	90,6	$84,8 \pm 0,3$	$84,6 \pm 0,5$
EMQ	93,6	74,8	59,0	96,0	89,5	91,3	95,6	87,2	$85,9 \pm 0,3$	$85,4 \pm 0,5$

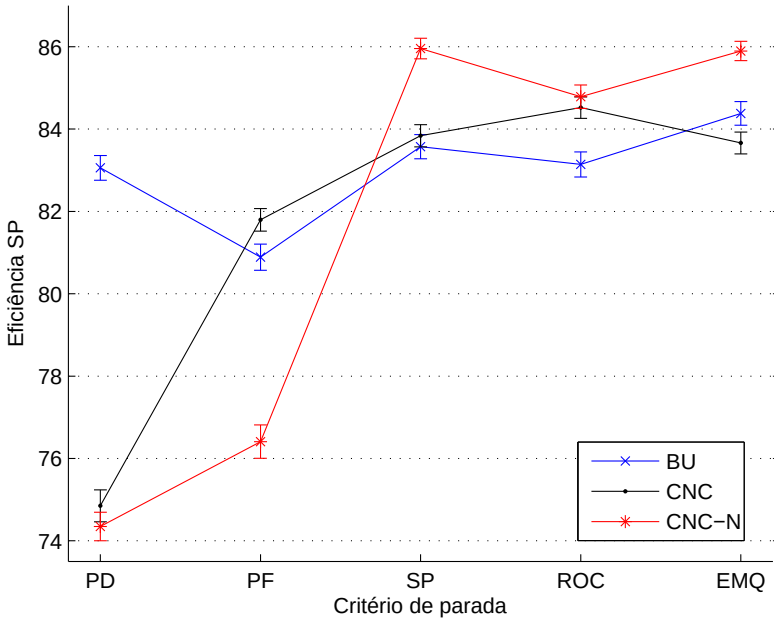
O melhor resultado obtido para a função objetivo CNC-N e pela parada por eficiência SP sinalizam que é importante considerar um balanço entre a capacidade de detecção e rejeição das classes no treinamento de classificadores classe-especialistas. Nas simulações realizadas, a capacidade de rejeição de cada especialista mostrou-se especialmente relevante ao desempenho global do sistema, refletindo-se, principalmente, nas confusões de identificação das classes.

6.4.2 Definição da topologia

Para a definição do número de neurônios da camada intermediária das redes classe-especialistas foram produzidos classificadores de complexidade crescente, segundo a técnica PCD construtiva, e avaliados 5 critérios de seleção, cada um baseado no valor de um índice de desempenho: PD, PF, ROC, SP e EMQ. Para o treinamento destes classificadores, foram utilizadas a função objetivo CNC-N e a parada



(a)



(b)

Figura 6.5: Eficiência média (a) e SP (b) do sistema especialista por função objetivo e critério de parada.

foi definida pela eficiência SP. Os valores dos índices obtidos para cada classe são exibidos nas Figuras 6.6 (PD e 100-PF), 6.7 (SP e ROC) e 6.8 (EMQ). Como, pelos índices PF e EMQ, a escolha é baseada no valor mínimo, optou-se, para facilitar a escolha, por esboçar, para os gráficos associados à PF, os valores correspondentes a 100-PF, e para aqueles relacionados à EMQ, o negativo dos valores, a fim de que, para todos os índices, a inspeção realizada buscasse o valor máximo.

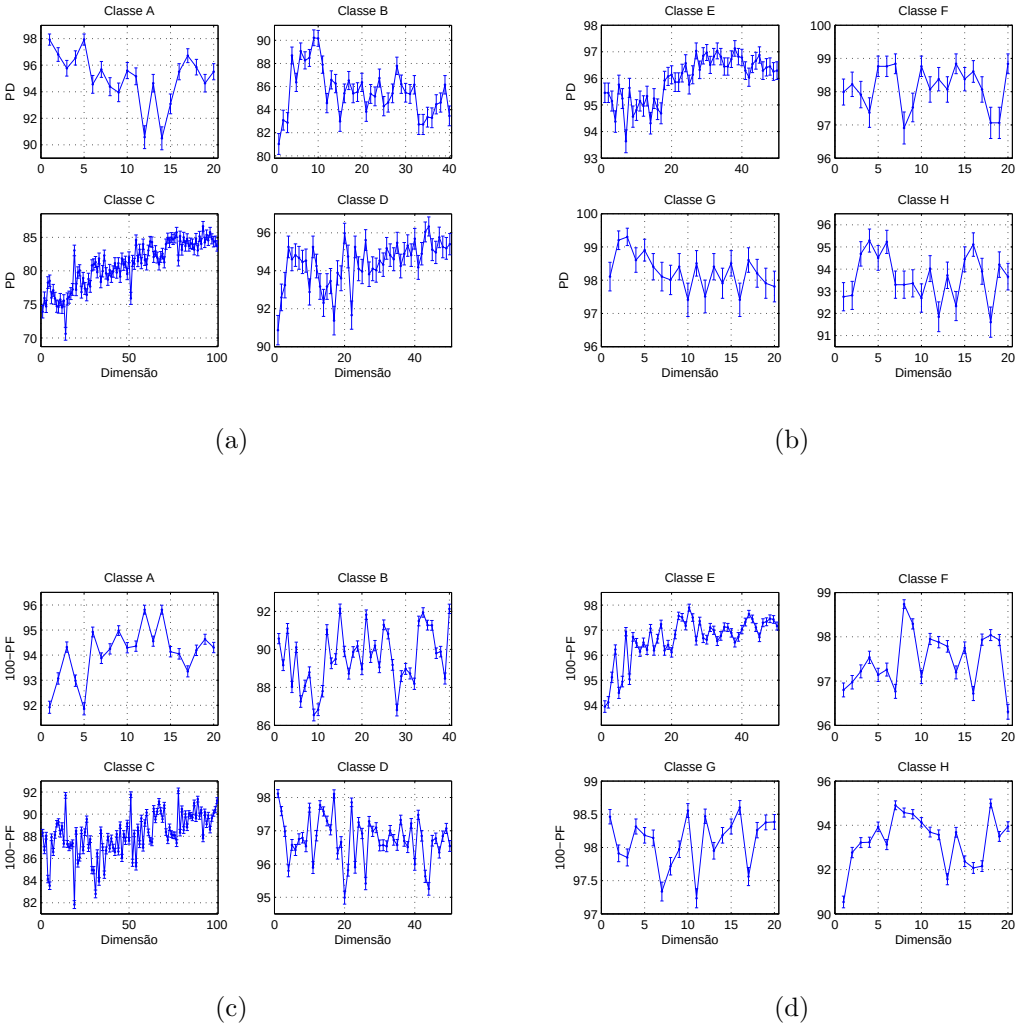


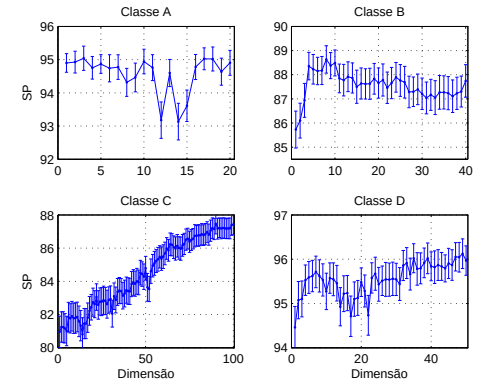
Figura 6.6: Eficiências de detecção e falso-alarme, por classe, para classificadores treinados segundo a técnica PCD (Veja o texto).

Em relação aos índices PD e PF, verifica-se uma variação expressiva dos valores, mesmo para topologias com um número similar de neurônios, como, por exemplo, para a classe A, considerando a detecção (PD) e topologias com 12 e 13 neurônios. Não é verificado, ainda, um comportamento monotônico dos índices com

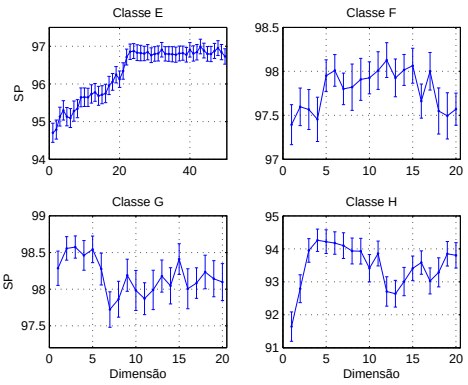
respeito à complexidade dos especialistas. Este fatores sinalizam que as flutuações inerentes ao processo de treinamento das redes possuem impacto significativo no desempenho inferido por estes índices. Em relação ao índice SP, tem-se um comportamento melhor, onde, para maioria das classes, excluindo a classe A, pode-se perceber uma tendência de crescimento, a qual se estabiliza posteriormente (classes C, D e E), ou ainda, pode estabilizar-se, temporariamente, e voltar a decrescer (classes B, F, G e H). Este comportamento indica que, a medida que cresce a topologia, tende a ocorrer uma melhora no desempenho dos classificadores, tanto em termos da detecção quanto na rejeição das classes. Como diferentes valores quanto à detecção e o falso-alarme podem resultar num mesmo desempenho quanto à eficiência SP, é compreensível a ocorrência de maiores flutuações para os índices PD e PF, sinalizando que parte expressiva das flutuações observadas está relacionada aos diferentes compromissos estabelecidos por cada treinemento entre a detecção e rejeição de classes, que resultam, no entanto, em classificadores equivalentes do ponto de vista da eficiência SP. Em relação a ROC, visto considerar, conjuntamente, as eficiências PD e PF, características similares à eficiência SP são observadas, porém algumas classes, tais como a A, F e G, apresentam flutuações demasiado pequenas, na faixa de $0,002$ e $0,004$. Por fim, para o erro quadrático médio, apenas as classes C e E apresentam uma curva com tendência crescente, com problemas similares aos identificados para os índices PD e PF.

Para cada índice foi identificada a topologia de menor complexidade e melhor desempenho. Na Tabela 6.2 é resumido o número de neurônios selecionado para as redes classe-especialistas de cada classe. Pode-se observar que todos os índices resultaram em sistemas de classificação cujo o número total de neurônios possui uma mesma ordem de grandeza, sendo sugeridas topologias com um total de 148 (PF) a 184 (ROC) neurônios.

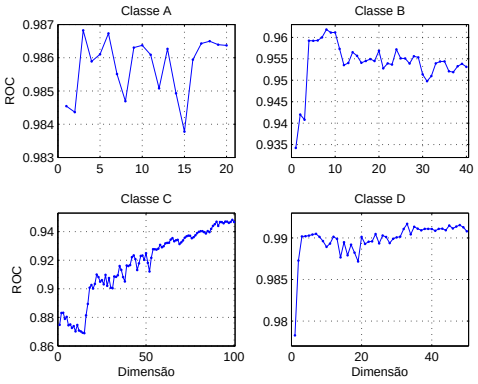
De posse das topologias identificadas, foram treinados novos classificadores, visto que a rede PCD foi utilizada apenas para a definição da topologia, e produzido um sistema classificador baseado no critério de integração por máximo. As eficiências classe-a-classe obtidas para cada critério de seleção são apresentadas na Tabela 6.3. É possível observar que, para cada classe, o melhor desempenho é obtido por um índice particular: PD, para as classes A e D; PF, para a classe E; SP, para a classe



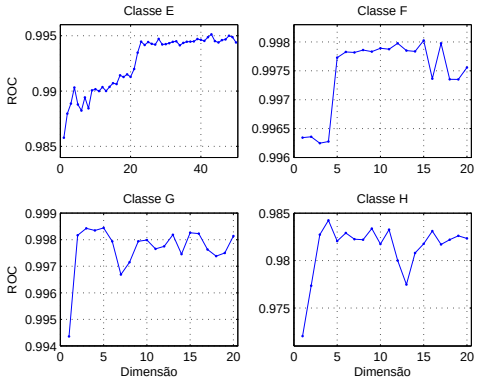
(a)



(b)

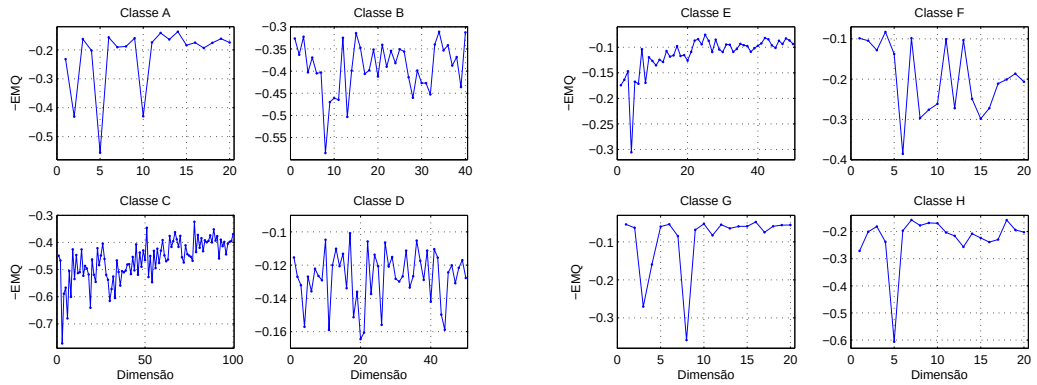


(c)



(d)

Figura 6.7: Eficiência SP e valor da área da ROC, por classe, para classificadores treinados segundo a técnica PCD (Veja o texto).



(a) (b)

Figura 6.8: Valor do erro quadrático médio, por classe, para classificadores treinados segundo a técnica PCD (Veja o texto).

C; ROC, para as classes B e H, e EMQ, para as classes F e G. Pode-se, verificar, no entanto, que do ponto de vista das eficiências média e SP, o índice de melhor desempenho é a área da ROC. Em relação aos índices SP e EMQ, seu desempenho mostrou-se inferior aos demais, em razão, principalmente, da deteção ruim verificada para a classe A. Através das matrizes de confusão, pode-se verificar que o principal responsável por este problema foi o especialista da classe C, o qual apresentou uma elevada taxa de falso alarme com respeito à classe A, ou seja, classificou número expressivo de eventos da classe A como provenientes da classe C.

Tabela 6.2: Topologias selecionadas por diferentes índices de desempenho (Veja o texto).

Critério	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Total**
PD	1	9	92	20	27	5	3	4	161
PF	12	15	78	1	25	8	1	8	148
SP	3	8	90	35	23	12	2	4	177
ROC	3	8	78	33	43	12	3	4	184
EMQ	12	15	99	16	25	4	1	7	179

Tabela 6.3: Eficiências classe-a-classe (%), média (%) e SP (%) para os sistemas de classificação baseados em redes classe-especialistas, com topologias selecionadas através de diferentes índices (Veja o texto).

Critério	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Eficiência média**	**Eficiência SP**
PD	**89,2**	72,7	68,2	**94,6**	94,6	90,2	96,5	89,5	$86,9 \pm 0,3$	$86,6 \pm 0,4$
PF	84,2	72,7	69,6	90,5	**94,7**	90,9	97,4	88,3	$86,0 \pm 0,3$	$85,7 \pm 0,5$
SP	69,3	73,2	**74,9**	92,1	91,8	88,6	96,6	90,5	$84,6 \pm 0,3$	$84,3 \pm 0,5$
ROC	87,6	**75,9**	70,2	94,0	94,0	91,9	96,7	**92,2**	$\mathbf{87,8 \pm 0,3}$	$\mathbf{87,6 \pm 0,4}$
EMQ	65,2	73,2	72,0	90,0	94,0	**92,5**	**97,6**	86,2	$83,8 \pm 0,3$	$83,4 \pm 0,6$

6.4.3 Integração

Os quatro critérios de integração considerados nesta análise foram: máximo, linear um, linear dois e neural, e utilizaram redes classe-especialistas com topologia definida pelo valor da área da ROC, treinadas segundo a função objetivo CNC-N, e cuja parada do treinamento foi realizada pela eficiência SP.

Para a obtenção dos parâmetros envolvidos no critério linear 1 foi realizado o cálculo da matriz pseudo-inversa dos dados; para o critério linear 2 foi produzida uma rede neural de camada única, com 8 neurônios lineares. O critério neural considerou uma rede integradora com 2 camadas, cujo número de neurônios na camada intermediária foi definido pela técnica PCD, utilizando, como índice de desempenho, a eficiência SP, o que resultou numa rede com 184 nós de entrada, 26 neurônios na camada intermediária e 8 neurônios de saída. Nesta rede integradora foi explorada uma codificação maximamente esparsa das classes [4], e todos neurônios possuem a tangente hiperbólica como função de ativação. Em razão do problema de mínimos locais [4], um total de 10 redes integradoras foram treinadas, segundo diferentes inicializações, sendo selecionada a de maior eficiência SP.

Na Tabela 6.4 são apresentadas as eficiências classe-a-classe, médias e SP dos sistemas de classificação classe-especialistas produzidos pelos diferentes critérios de integração. Considerando cada classe individualmente, há um melhor desempenho do critério de máximo para as classes D, G e H; do linear 1, para as classes C, D,

E e F; do linear 2, para a classe G, e do neural, para as classes A e B. O melhor desempenho global, segundo as eficiências média e SP, é pelo critério neural, em razão do desempenho obtido para as classes A e B. Em segundo lugar, tem-se o critério de máximo, bastante atrativo por não exigir o ajuste de parâmetros e por sua simplicidade, com desempenho inferior em 0,9 pontos percentuais para ambas eficiências em relação ao primeiro. Na comparação entre os critérios máximo e neural, há um melhor desempenho do primeiro para as classes D, G e H.

Tabela 6.4: Eficiencias classe-a-classe (%), média (%) e SP (%) para os sistemas de classificação classe-especialistas baseados em diferentes critérios de integração.

Integração	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Eficiência média**	**Eficiência SP**
Máximo	87,6	75,9	70,2	**94,0**	94,0	91,9	**96,7**	**92,2**	$87,8 \pm 0,3$	$87,6 \pm 0,4$
Linear 1	80,1	69,0	**73,9**	**94,0**	**94,7**	**93,4**	95,0	86,0	$85,8 \pm 0,3$	$85,4 \pm 0,5$
Linear 2	82,2	63,7	73,0	93,9	94,6	90,5	**96,7**	88,1	$85,3 \pm 0,3$	$85,0 \pm 0,5$
Neural	**88,0**	**81,6**	72,1	93,7	95,3	92,5	95,4	90,8	$\mathbf{88,7 \pm 0,2}$	$\mathbf{88,5 \pm 0,4}$

6.5 Detecção e inclusão de novas classes

No ambiente de sonar é provável que o sistema de classificação automática se depare com diferentes condições operativas, navios, ou mesmo, novas classes. Tendo em vista as restrições quanto à caracterização estatística das classes existentes no conjunto de dados utilizado neste trabalho, é de especial importância que o sistema de classificação proposto possua mecanismos para a identificação de cenários desconhecidos. Esta identificação aumenta a confiabilidade das informações providas pelo sistema, e pode ser utilizada como um alerta da necessidade de maior atenção do operador em relação ao contato. Outra possibilidade é que os casos identificados como desconhecidos sejam incorporados à base de dados ou utilizados para o retreino do sistema.

A detecção de novas classes é um problema, normalmente, desafiador. Na literatura há várias propostas, sendo a escolha da técnica mais apropriada dependente

da aplicação [234]. Um requisito para todas as técnicas é estabelecer um compromisso entre a identificação de eventos novos e a detecção de eventos conhecidos [234], ou seja, o sistema não deve confundir informação generalizada como nova [235]. É desejável ainda que o desempenho na detecção de novos eventos seja independente do número de características e classes envolvidas, mostrando-se robusto em conjuntos de dados desbalanceados, com pequeno número de amostras e nível significativo de ruído. Prover mecanismos de adaptação do classificador aos eventos identificados como novos é também útil [236]. Para aplicações que operam em tempo real, faz-se necessário ainda restringir o custo computacional das técnicas utilizadas. Dois enfoques principais para a detecção de novos eventos são verificados na literatura: um baseado em técnicas estatísticas [234]; e outro, em técnicas neurais [237].

Em linhas gerais, nas técnicas estatísticas é realizada uma modelagem das características estatísticas dos dados. De posse de um evento é inferida a probabilidade do dado em pertencer aos modelos produzidos. Caso esta probabilidade seja inferior a um limiar, o evento é considerado como novo. Em aplicações com múltiplas classes, um limiar pode ser estabelecido para cada classe [238]. Quanto à modelagem, podem ser utilizadas técnicas paramétricas e não-paramétricas [239]. Pela primeira, é pressuposto que os dados são provenientes de distribuições conhecidas, cujos parâmetros são ajustados com base nos dados, o que demanda um extensivo conhecimento a priori do problema. Nos modelos não-paramétricos, tanto a densidade quanto seus parâmetros são derivados dos dados.

Nos critérios baseados em estimação paramétrica, o modelo gaussiano de misturas [49] é freqüentemente utilizado, como em [240], no diagnóstico de casos de epilepsia. Em [241], para a detecção de massas em mamogramas, é proposta a partição do espaço dos dados pelo algoritmo *k-means* [49], sendo utilizado um modelo gaussiano de mistura e limiar para cada partição. Um problema dos métodos paramétricos é o curso da dimensionalidade [234], isto é, o fato que o número de eventos necessários para uma apropriada estimação dos modelos cresce exponencialmente com a dimensão dos dados. Assim, freqüentemente, estas técnicas são utilizadas em problemas com dimensão reduzida, ou associados a técnicas de compactação e/ou extração de características.

Entre as técnicas não-paramétricas são comuns a estimação das distribuições

através do método de Parzen [49], ou a produção de agrupamentos com grupos esféricos ou elípticos, para os quais o algoritmo *k-means* é, freqüentemente, utilizado visando definir o centro dos grupos. No último caso, um evento é classificado como novo caso se situe na região exterior dos grupos, avaliação que, usualmente, utiliza a distância euclidiana, para agrupamentos esféricos, ou de Mahalanobis [49], para os agrupamentos elípticos. Alguns trabalhos nesta linha são [242], na detecção de falhas de motores a jato; e [243], que visa identificar falhas de rotores.

Um critério baseado na idéia de instabilidade de classificação é proposto em por [235]. Segundo este critério, num grupo de classificadores destinados a uma mesma tarefa, as divergências de classificação podem ser utilizadas como um indício de eventos desconhecidos. O artigo propõe a produção de múltiplas amostras *bootstrap* dos dados, gerando-se, para cada amostra, um classificador baseado no critério de Fisher [1]. Um evento é identificado como novo caso a variância da saída dos classificadores seja superior a um limiar arbitrário.

Redes neurais são classificadores especialmente interessantes para problemas complexos, com dados de alta-dimensionalidade e sujeitos a restrições quanto à caracterização estatística. A detecção de novos eventos em redes neurais é, normalmente, mais crítica que para os modelos estatísticos, tendo em vista que se tratam de discriminadores e não-detectores, realizando, em geral, a separação dos alvos através de regiões de decisão abertas definidas por hiperplanos [237]. Assim, alguns trabalhos [226, 227, 244, 245, 246] propõem a produção de novas amostras no conjunto de treinamento visando melhor definir as fronteiras das classes, isto é, prover seu encapsulamento. Uma estratégia comum consiste em produzir eventos espacialmente distribuídos ao redor das classes conhecidas. Algumas propostas são: Zhang [244], que utiliza algoritmos genéticos; Wei [246], o qual realiza a alteração aleatória das variáveis dos dados, e as referências [226, 227, 245] que exploram distribuições gaussianas. Estas estratégias sofrem, no entanto, do problema do curso da dimensionalidade, logo são apenas aplicáveis a problemas de dimensionalidade reduzida.

Para classificadores MLP [4], a detecção de novos eventos é, usualmente, realizada através da comparação da saída rede com um limiar. Em [247], numa aplicação de reconhecimento de caracteres manuscritos, é proposto um critério que

identifica como conhecidos os eventos que satisfizerem 3 condições: (1) o valor da saída do neurônio vencedor for superior a um limiar L_1; (2) a segunda maior saída for maior que L_2, e (3) a diferença entre as saídas destes neurônios for superior a L_3. Em [248] é proposto um critério similar, contemplando, no entanto, apenas as condições (1) e (2). Outros trabalhos que exploram esta estratégia são [226, 227, 245]. Em [249] é proposto que o limiar seja obtido através da otimização de uma função que atribuiu custos para a taxa de reconhecimento, rejeição e classificação errônea dos eventos. No trabalho [2], que envolve a classificação de sinais de sonar provenientes de 4 classes de navios, é proposta a identificação de novas classes através de limiares, escolhidos através de um compromisso previamente estabelecido entre a detecção de classes desconhecidas e o reconhecimento de novas classes.

Em relação a classificadores baseados em redes ART [56, 205], a detecção de novos eventos pode ser baseada em sua pertinência aos grupos definidos pelos neurônios. Entre trabalhos tem-se: [250], no reconhecimento de imagens de radar, e o já descrito [2]. Há propostas baseadas em redes de base radial [4], as quais seguem a idéia de prover modelos probabilísticos para os dados, como em [251], no reconhecimento de locutores. Neste trabalho, a identificação de novos eventos é baseada na distância de Mahalanobis [49] ao *kernel* de maior resposta. Outra linha são as redes auto-associativas [4], que são treinadas para reproduzir, na saída, o vetor de entrada da rede. A identificação de novos eventos é baseada no erro cometido na saída da rede, que é comparado com um limiar. Alguns trabalhos nesta linha são [252, 253, 254]. Por fim, utilizando mapas auto-organizáveis [49], uma aplicação é o diagnóstico de falhas em [255], e a detecção de invasão de redes de computadores, realizada em [256].

Tão relevante quanto identificar novos cenários é permitir que o classificador se adapte a eles. A adaptação é, normalmente, um processo complexo, o qual deve ser realizado cuidadosamente, a fim de evitar uma degradação expressiva de desempenho para as classes conhecidas. Em outras palavras, deve-se evitar perdas do conhecimento adquirido, em especial, para aplicações críticas com fortes requisitos quanto a segurança, tais como o sistema de classificação proposto.

Neste contexto é, portanto, contra-indicado que a inclusão de novos cenários demande o projeto de um classificador completamente novo, sem utilizar ne-

nhum conhecimento armazenado no classificador original. Em outras palavras, é contra-indicado o retreino do zero, o que torna desejável a adoção de classificadores construtivos. Em [257] são discutidos alguns critérios para permitir a incorporação de conhecimento a classificadores MLP. A proposta é manter "congelados"os pesos da rede original, os quais são supostos úteis na modelagem de parte da função objetivo, realizando-se a inserção e o treinamento de novos neurônios. Uma alternativa à inserção de novas classes, já discutida anteriormente, são as redes neurais classe-especialistas, cuja inclusão de classes é realizada pela adição de novos módulos classe-especialistas, seguida por uma eventual alteração do critério de integração.

Incorporar uma nova classe de contato exige a existência de um conjunto de aquisições, onde as diferentes características desta classe sejam, apropriadamente, caracterizadas, em especial, aquelas responsáveis por discriminá-la das classes conhecidas. Uma primeira possibilidade é que esta incorporação seja realizada fora da operação real do sistema (*off-line*). Neste caso, de posse de um conjunto caracterísro para a nova classe, realiza-se o retreinamento do sistema. Outra possibilidade, mais problemática e de implementação mais complexa, é realizar a adaptação ao longo da operação do sistema.

Na adaptação fora de operação, um caminho possível para a caracterização das novas classes é prover o sistema de classificação com um mecanismo de detecção de novos eventos, os quais, uma vez identificados, são armazenados num banco de dados específico. Este banco, posteriomente analisado e tratado, pode constituir a base de retreino do sistema. Cabe notar que este conjunto pode possuir eventos de uma ou mais classes novas, assim como conter eventos espúrios de classes conhecidas. Faz-se necessário, portanto, identificar quais eventos são espúrios, quantas novas classes existem, e quais eventos as constituem. Raros são os trabalhos que discutem este problema. Em [227] é proposta que a separação entre classes novas e eventos espúrios seja realizada de forma manual, com base na análise de um agrupamento produzido sobre o banco de dados de retreino. Para a produção deste agrupamento, é proposta a utilização do algoritmo *k-means*, com número de grupos definidos pelo índice de *Davies-e-bouldin* [160], ou através de mapas auto-organizáveis [49]. Quanto à identificação de grupos como relacionados à eventos espúrios ou novas classes, é sugerida a avaliação da distância de Bhattacharya [49].

Com respeito à detecção e inclusão de classes, a seguir serão avaliados diferentes critérios aplicáveis a classificadores classe-especialistas, assim como será discutida a inclusão, fora de operação, de novas classes ao sistema.

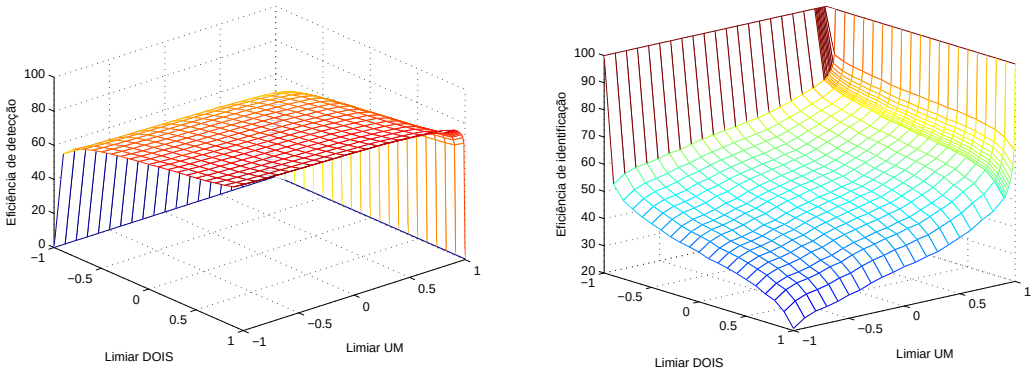
6.5.1 Critérios para a identificação de novas classes

Um critério simples para a identificação de novas classes, aplicável a sistemas de classificação baseados em redes neurais classe-especialistas, é considerar limiares para a detecção de cada classe, ou seja, um evento só é classificado como proveniente de dada classe se a saída do especialista a ela associado for superior a um limiar arbitrário. Para uma maior simplicidade, é conveniente considerar um mesmo limiar para todas as classes. Este critério de decisão é baseado no pressuposto que, para eventos de classe desconhecidas, há uma tendência dos especialistas em apresentar baixos valores de saída, ou seja, de rejeitá-los. Em outras palavras, como a saída dos classificadores é probabilística, eventos associados a uma baixa probabilidade de pertencer às classes são considerados como provenientes de uma nova classe. Cabe observar, no entanto, que esta tendência é tanto maior quanto melhor for o encapsulamento de cada classe, logo está relacionada às regiões de decisão construídas por cada especialista, que, por sua vez, são dependente das características das classes utilizadas e da sistemática adotada no treinamento.

Um critério mais robusto pode ser obtido se, além de atendida a exigência do critério anterior, exigir-se que o segundo maior valor de saída dos especialistas seja inferior a um segundo limiar arbitrário (limiar dois). Este critério explora a idéia que, para eventos de classes conhecidas, há uma tendência das saídas dos classificadores em apresentar uma maior coerência, ou seja, em apresentar apenas uma saída com valor alto, preferencialmente, próxima de +1; enquanto as demais, nas vizinhanças de -1.

Em ambos critérios, cada limiar estabelece um compromisso entre o reconhecimento de classes conhecidas e a identificação de novas classes. Para fins ilustrativos, na Figura 6.9 são apresentadas a capacidade de detecção e de identificação de um sistema especialista arbitrário, considerando diferentes escolhas para os dois limiares. Pode-se perceber que, suposto fixo o limiar um, ao reduzir-se o limiar dois de +1 até -1, é verificada uma redução na detecção e aumento na identificação. Fato

similar ocorre se suposto o limiar dois fixo, e o limiar um variar, de forma crescente, na faixa de -1 a +1.



(a) (b)

Figura 6.9: Eficiências de detecção de classes conhecidas (a) e de identificação de novas classes (b) por escolha do limiares de decisão (um e dois).

Um gráfico que permite uma melhor avaliação das habilidades de um critério de identificação arbitrário quanto à detecção de classes conhecidas e o reconhecimento de novas classes é apresentado na Figura 6.10. Neste gráfico, o valor da eficiência de detecção é esboçado em função da eficiência de identificação, e cada curva corresponde a uma escolha do limiar dois. Pode-se perceber que o ganho na identificação é sempre traduzido em perda de detecção, em especial, ao reduzir-se o limiar dois.

Uma alternativa mais sofisticada para a identificação de novas classes consiste em produzir agrupamentos sobre as saídas dos especialistas. Neste caso, um evento é classificado como proveniente de classe conhecida caso o vetor definido pelas saídas dos especialistas pertença a um dos grupos definidos pelo agrupamento. Na Figura 6.11 este critério é ilustrado para um sistema especialista, arbitrário, de 2 classes. Pode-se perceber que dois grupos esféricos, com centros $(-1, +1)$ e $(+1, -1)$, foram definidos. Estes centros correspondem a uma composição dos alvos utilizados no treinamento dos especialistas 1 e 2. Neste caso, um evento é classificado como proveniente de dada classe, apenas se o conjunto das saídas dos especialistas definirem um ponto no interior do círculo associado a esta classe.

No exemplo anterior, através da escolha do raio, grupos de diferentes dimen-

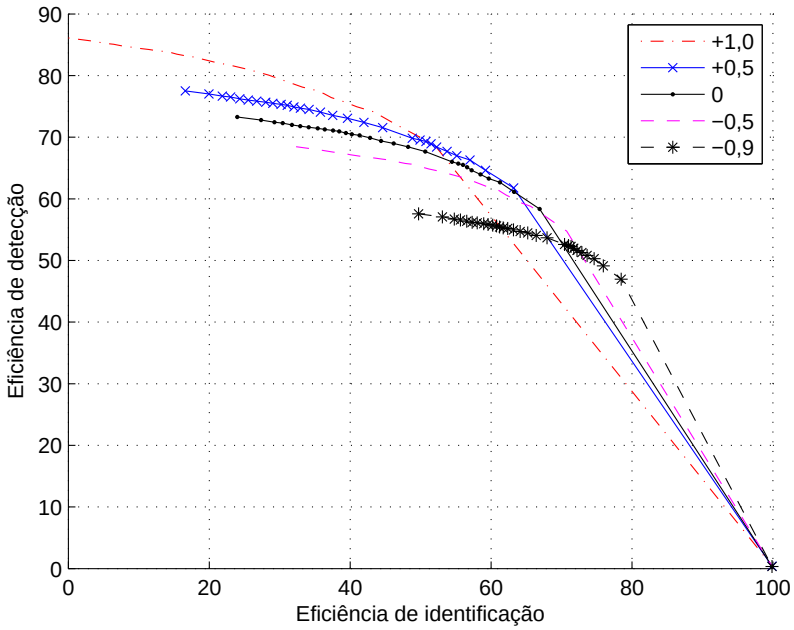


Figura 6.10: Curvas de detecção de classes conhecidas e identificação de novas classes para diferentes escolhas do limiar dois

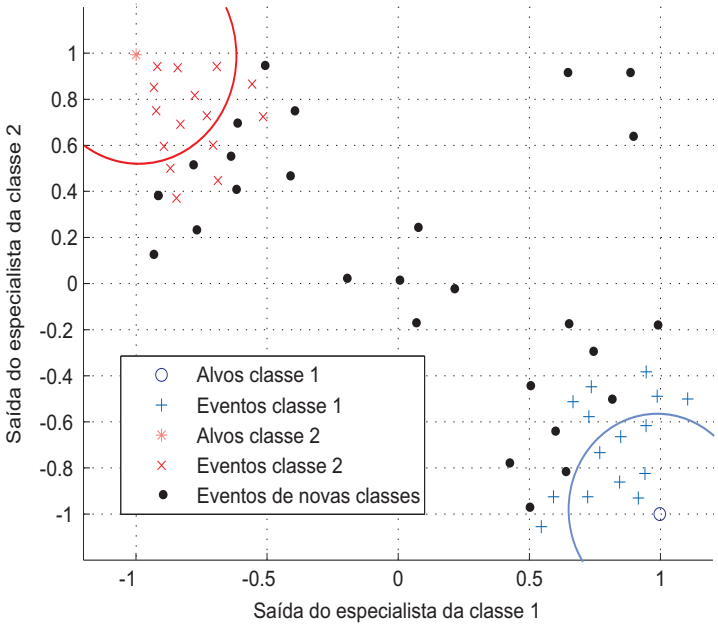


Figura 6.11: Ilustração do critério de identificação de novas classes através de agrupamento para um sistema especialista de duas classes (Veja o texto).

sões podem ser definidos. Tanto maior o tamanho do grupo, maior é a detecção, porém menor é a identificação. Assim, a escolha do tamanho do grupo define um compromisso entre as capacidades de detecção e identificação do sistema. De forma similar aos critérios baseados em limiar, este critério é também depende da qualidade do encapsulamento das classes provido de cada especialista.

Para um número arbitrário K de classificadores classe-especialistas, um critério simples consiste em definir uma hiperesfera, de centro $\mathbf{c}^i$ $(1 \leq i \leq K)$, e mesmo valor de raio r, para maior simplicidade, por classe. Um evento será identificado como proveniente de uma classe conhecida se:

$$\exists i \quad | \quad (y_1 - c_1^i)^2 + \ldots + (y_K - c_K^i)^2 \leq r, \tag{6.22}$$

para y_l correspondente à saída do l-ésimo especialista $(1 \leq l \leq K)$, e c_j^i definido pela j-ésima coordenada do centro da hiperesfera associada à detecção da i-ésima classe. Em outras palavras, caso a distância euclidiana do vetor definido pelas saídas dos especialistas em relação a centro arbitrário $\mathbf{c}^i$ seja inferior a um limiar r, o evento é considerado como proveniente da i-ésima classe.

Um critério alternativo consiste em utilizar a distância de Mahalanobis [49], para a qual o critério anterior resulta em:

$$\exists i \quad | \quad (\mathbf{y} - \mathbf{c}_i)^T \mathbf{R}_{\mathbf{y}_i}^{-1} (\mathbf{y} - \mathbf{c}_i) \leq r, \tag{6.23}$$

onde $\mathbf{y}$ é um vetor formado pelas saídas dos especialistas para um evento arbitrário, e $\mathbf{R}_{\mathbf{y}_i}$ é a matriz de covariância das saídas dos especialistas para os eventos da i-ésima classe. A distância de Mahalanobis define grupos por hiperelipsóides, com volume relacionado ao valor de r, que possuem eixos definidos pelos autovetores da matriz $\mathbf{R}_{\mathbf{y}_i}$.

Quanto à escolha dos centros dos grupos, duas possibilidades foram analisadas: a primeira, com centros definidos por uma composição de vetores alvos, conforme ilustrado no exemplo anterior; e uma segunda, onde $\mathbf{c}_i = E[\mathbf{y}_i]$, ou seja, o i-ésimo centro correspondente ao centróide das saídas dos especialistas para dados da i-ésima classe.

6.5.2 Inclusão de novas classes

A inclusão de novas classes aqui discutida considerará uma atualização fora de operação do sistema, para a qual é suposta a existência de conjuntos de dados que caracterizem, apropriadamente, as classes a serem incluídas. Neste caso, o processo de inclusão é simples, consistindo na produção e inserção de classificadores para as novas classes, com uma eventual adaptação do critério de integração, se necessário.

Na inserção de novos especialistas ao sistema classificador, é possível identificar dois grupos de módulos: os originais, treinados segundo um número mais restrito de classes; e os novos, cujo treinamento contemplou todas as classes. É possível, no entanto, que alguns dos módulos originais apresentem um desempenho não satisfatório para as novas classes, o que pode ser identificado através da matriz de confusão [4] do sistema classificador. Para estes casos, pode-se optar pelo retreinamento dos especialistas de desempenho mais crítico.

6.6 Resultados para a identificação e inclusão de novas classes

Para a avaliar a identificação e a inclusão de novas classes ao sistema especialista, as classes A, B, C e H foram supostas conhecidas, e as demais (D, E, F e G) foram consideradas novas. Esta escolha foi baseada nas análises desenvolvidas no Capítulo 5, que identificaram o primeiro conjunto como classes de classificação mais crítica.

De forma coerente aos resultados anteriores, o treinamento dos especialistas considerou a função objetivo CNC-N, a eficiência SP como critério de parada, e o máximo foi utilizado para a integração. A escolha do número de neurônios seguiu procedimento similar ao realizado na seção 6.4.2. Na Tabela 6.5 são apresentadas as topologias sugeridas por cada índice. Os valores de eficiência classe-a-classe, média e SP dos sistemas de classificação obtidos são apresentados na Tabela 6.6.

Pode-se verificar que são propostas topologias contendo um total de 38 (ROC) a 77 (EMQ) neurônios. O desempenho dos diferentes índices é razoavelmente equilibrado, destacando-se o SP e o EMQ, em razão das eficiências obtidas para as classes A e B. Na comparação SP e EMQ, o primeiro é melhor, já que utiliza classificado-

Tabela 6.5: Topologias dos classificadores classe-especialistas selecionadas por índice de desempenho.

	A	**B**	**C**	**H**	**Total**
PD	3	5	30	5	43
PF	8	6	23	8	45
SP	4	5	48	8	65
ROC	3	5	5	25	38
EMQ	11	10	48	8	77

Tabela 6.6: Eficiências classe-a-classe (%), média (%) e SP (%) considerando a seleção das topologias dos classificadores classe-especialistas segundo diferentes índices de desempenho.

Critério	**A**	**B**	**C**	**H**	**Eficiência Média**	**Eficiência SP**
PD	86,1	83,2	**75,8**	**94,7**	$84,8 \pm 0,7$	$85,0 \pm 0,4$
PF	87,7	82,5	73,9	94,2	$84,4 \pm 0,7$	$84,6 \pm 0,4$
SP	**91,3**	**85,7**	73,1	94,3	$\mathbf{85,9 \pm 0,6}$	$\mathbf{86,1 \pm 0,3}$
ROC	91,0	82,3	68,6	94,0	$83,7 \pm 0,7$	$84,0 \pm 0,4$
EMQ	91,2	84,1	74,1	94,4	$85,8 \pm 0,6$	$86,0 \pm 0,3$

res menos complexos e resulta em eficiências média e SP ligeiramente superiores ao EMQ. É interessante observar que, novamente, o índice de melhor desempenho leva em consideração, conjuntamente, a capacidade de detecção e rejeição de classe dos especialistas.

6.6.1 Identificação de novas classes

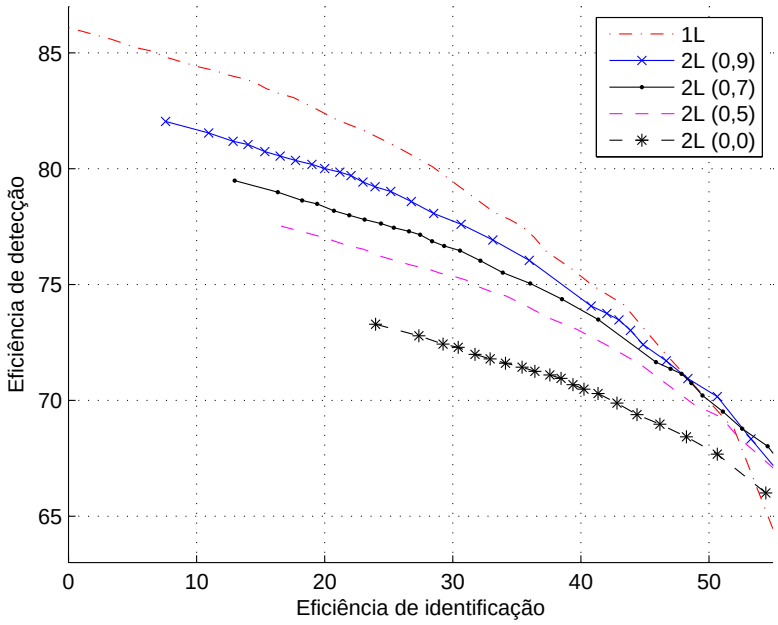
Para simplificar a análise, a comparação entre as diferentes propostas foi baseada no valor médio das eficiências de detecção e identificação das classes, utilizando curvas onde a primeira é expressa como função da última. Na Figura 6.12 são apresentadas as curvas obtidas para os critérios baseados em limiar único (1 L) e em 2 limiares (2L), o último considerando diferentes valores $(0,9; 0,7; 0,5; 0,0)$ para o limiar dois, segundo eficiências de identificação na faixa de 1% a 54% (a) e de 55% a 100% (b).

Considerando valores de identificação de até $46,4\%$, um conjunto de maiores eficiências de detecção é obtido para o critério baseado em limiar único, situando-se entre $86,1\%$ e $72,2\%$.

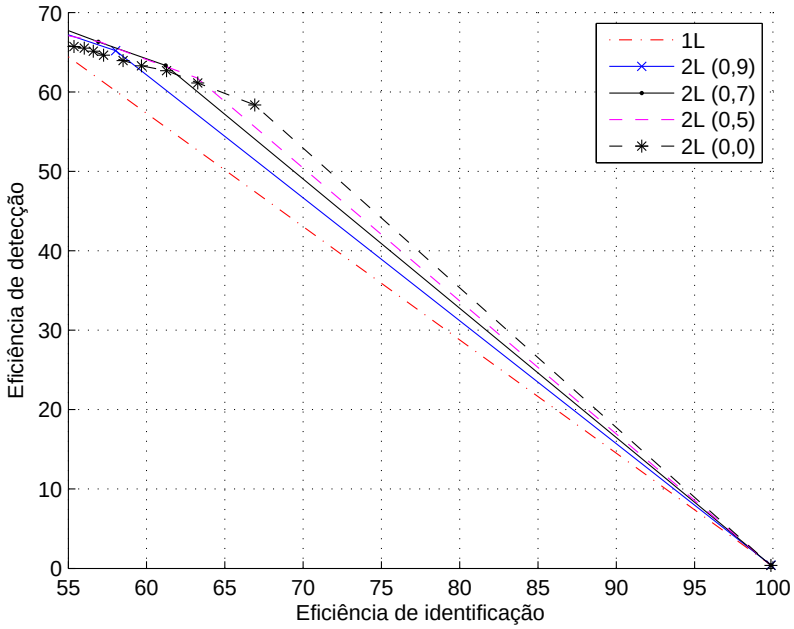
Uma segunda análise considerou, apenas, os critérios baseados em agrupamentos, utilizando uma hiperesfera ou hiperelipsóide por classe, com centro definido por uma composição dos alvos (Alvo), ou através do centróide das saídas dos especialistas (Centróide). A Figura 6.13 exibe os resultados obtidos para faixas de eficiências de identificação de 0% a 50% e de 50% a 100% (b).

Para uma eficiência de identificação de até $43,7\%$, o melhor desempenho é obtido utilizando hiperesferas com centros definidos pela composição dos alvos. Nesta faixa, as eficiências de detecção variam de $86,1\%$ a $68,6\%$. Uma comparação entre os critérios de limiar único e hiperesfera é realizada na Figura 6.14. Para eficiências de identificação de até $56,8\%$, o critério baseado em limiar único apresenta melhor desempenho.

Considerando o critério de limiar único, a Figura 6.15 apresenta, por cada classe, as eficiências de detecção e identificação como função do valor do limiar. Na Tabela 6.7 são resumidos alguns dos valores obtidos. É possível observar que a classe B é a que apresenta maiores variações da eficiência de detecção com respeito à escolha do limiar. Para a maioria das classes, exceto a B, a eficiência de detecção começa a

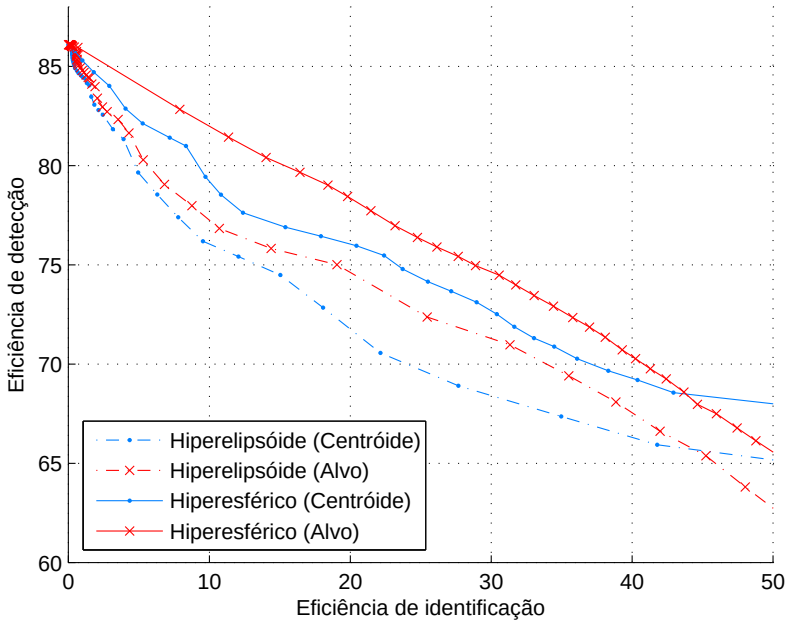


(a)

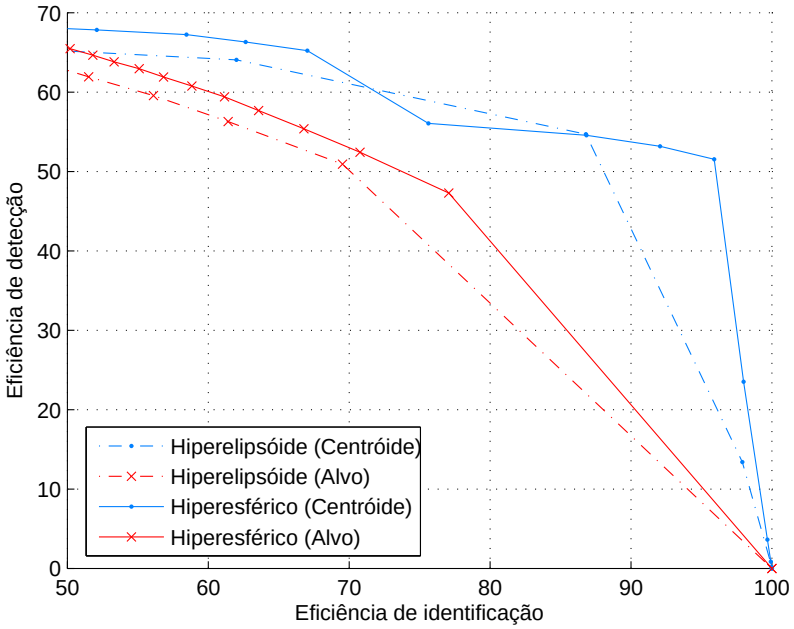


(b)

Figura 6.12: Eficiências de detecção como função das eficiências de identificação, para os critérios baseados em 1 ou 2 limiares (Veja o texto).



(a)



(b)

Figura 6.13: Eficiências de detecção como função das eficiências de identificação para os critérios baseados em agrupamento (Veja o texto).

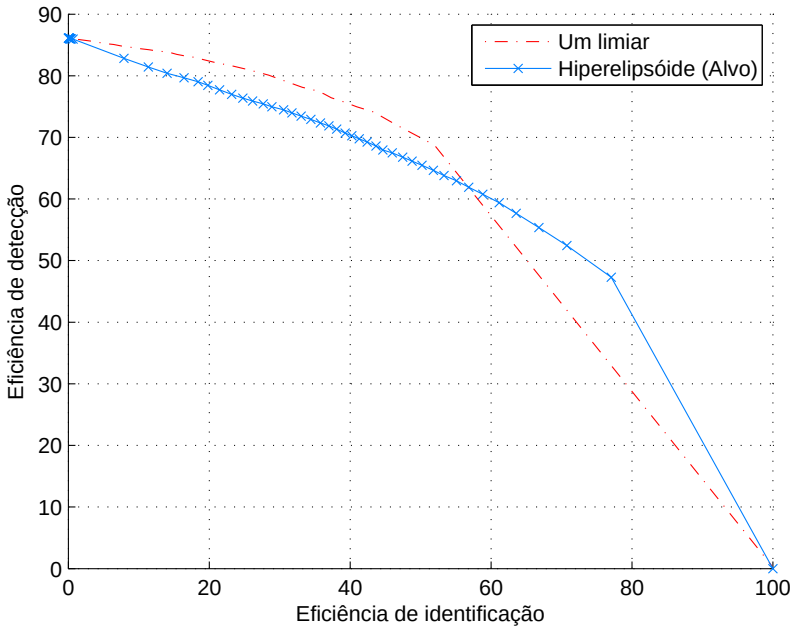


Figura 6.14: Comparação das eficiências de detecção como função das eficiências de identificação para os critérios limiar único e hiperesfera, com centro definido pela composição dos alvos.

reduzir, de forma considerável, apenas para valores de limiar superiores a 0,5. Em relação à identificação de novas classes, um melhor resultado é obtido para a classe G; em segundo lugar, para as classes A e B; e por fim, para a classe F. Verifica-se que as curvas de identificação tendem a apresentar uma maior taxa de crescimento na faixa de 0,5 a 1,0. Para um limiar de 0,95, por exemplo, é possível obter uma taxa de identificação mínima de 21% e máxima de $61,2\%$ pontos percentuais, para as classes F e G, respectivamente.

Dos resultados anteriores, verifica-se que o desempenho do sistema de detecção é razoavelmente dependente da nova classe considerada. Para investigar a quais classes conhecidas são, mais freqüentemente, atribuídos os eventos das novas classes foi elaborado o gráfico da Figura 6.16. Neste gráfico, para cada nova classe, é apresentado o percentual de eventos originais que são classificados como provenientes das classes A, B, C e H para diferentes valores de limiar. Em relação à classe D, uma maior confusão é obtida pelo especialista da C. Para a classe E, as principais confusões são cometidas pelos especialistas C e A. Em relação à classe F, há confusões

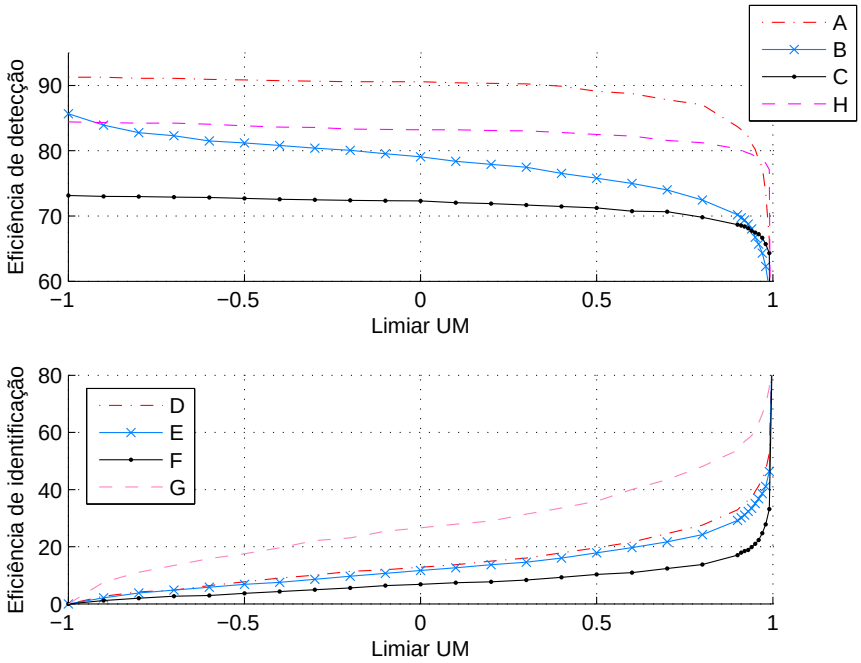


Figura 6.15: Eficiências de detecção e identificação por classe e valor de limiar.

Tabela 6.7: Eficiências de detecção e identificação classe-a-classe e média por valor de limiar

	Detecção					**Identificação**				
Limiar	**A**	**B**	**C**	**H**	**Valor médio**	**D**	**E**	**F**	**G**	**Valor médio**
-1	91,3	85,7	73,2	94,3	**86,1**	0	0	0	0	**0**
0,8	87,0	72,5	69,8	91,1	**80,1**	27,5	24,2	13,8	48,1	**28,4**
0,9	83,7	70,1	68,6	90,0	**78,1**	32,9	29,1	17,0	53,8	**33,2**
0,95	80,2	66,8	67,5	88,5	**75,8**	39,4	35,1	21,0	61,2	**39,2**

expressivas de todos classificadores, exceto o relacionado à classe B. Para a classe G, a confusão mais expressiva é devida ao especialista da classe B. Pode-se observar que, para a maior parte das novas classes, as maiores confusões estão relacionadas ao especialista da classe C. Estes resultados sinalizam, especialmente para a classe C, a necessidade de prover classificadores especialistas que melhor encapsulem as classes.

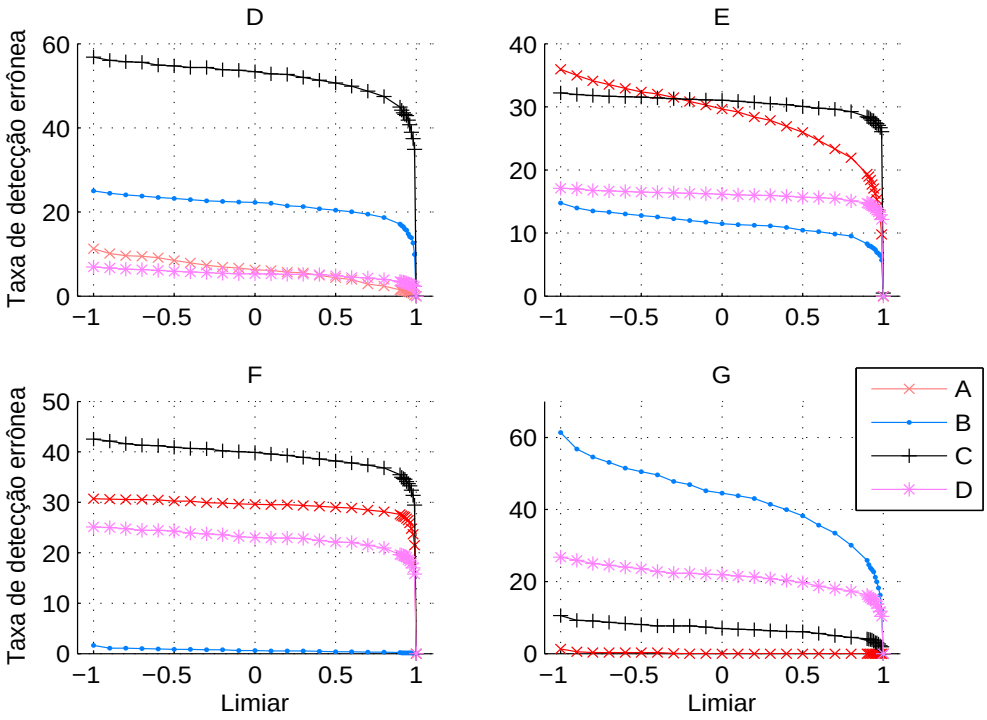


Figura 6.16: Percentual de eventos das classes novas identificados errôneamente como provenientes de classes conhecidas para diferentes escolhas de limiar.

Quanto à escolha do limiar, nas análises anteriores, é possível proceder a escolha com base num compromisso entre a detecção e a identificação das classes. Uma vez que a estimação da capacidade de identificação demanda o conhecimento a priori das novas classes, em cenários reais de operação, este critério não poderá ser utilizado. Uma alternativa é estabelecer uma degradação máxima admissível quanto à capacidade de detecção, para a qual o limiar é definido. Pelos dados da Tabela 6.7, por exemplo, se for arbitrado uma eficiência média mínima de 80%, o limiar escolhido seria $0,8$.

6.6.2 Inclusão de novas classes

A inclusão de novas classes considerou todas as classes D, E, F e G, tidas como novas, e seus classificadores classe-especialistas foram treinados segundo a função objetivo CNC-N, com parada por eficiência SP, e a topologia foi definida segundo a Tabela 6.2. A integração foi realizada pelo critério de máximo.

Na Tabela 6.8 são apresentados os valores das eficiências classe-a-classe e SP considerando o sistema classificador antes e após a inclusão das novas classes. Para a última situação, dois cenários foram considerados: com e sem o retreino dos classificadores originais. Em relação as classes A, B, C e H (originais) é verificada uma pequena redução da eficiência de detecção, em especial, para a classe B, mais marcante na ocorrência de retreino. Para as novas classes, ocorrendo o retreino, há um ganho expressivo, de, no mínimo, 15 (classe D) e, no máximo, 43 (classe F) pontos percentuais. Para a classe G, este ganho é bem menor, da ordem de 0,8 pontos percentuais.

Tabela 6.8: Eficiências classe-a-classe e SP na integração de novas classes realizada com ou sem o retreino dos classificadores originais.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Eficiência (SP)**
4 classes	91,3	85,7	73,1					94,3	$86,1 \pm 0,3$
Sem retreino	89,3	79,2	71,3	79,2	76,4	48,5	95,9	93,7	$78,5 \pm 0,7$
Retreino completo	87,6	75,9	70,2	94,0	94,0	91,9	96,7	92,2	$87,6 \pm 0,4$

A matriz de confusão associada ao sistema de classificação sem retreino é apresentada na Tabela 6.9. Pode-se verificar que, para os eventos da classe D, a maior confusão é realizada pelo especialista da classe C, segundo uma taxa de $15,9\%$. Para a classe E, tem-se confusões providas pelos especialistas das classes C e H, com $13,9\%$ e $5,4\%$, respectivamente. Para a classe F, tem-se taxas de confusão de $24,6\%$ (classe C), $14,3\%$ (classe H) e $11,7\%$ (classe A). É possível perceber que o especialista da classe C está associado à parcela significativa das confusões verificadas, tornando-o um forte candidato ao retreino. Entre outros candidatos, nesta ordem, tem-se os especialistas das classes H e A. Deste modo, foram consideradas diferentes possibilidades quanto ao retreino dos classificadores: uma primeira, envol-

Tabela 6.9: Matriz de confusão para o sistema classificador após integração das novas classes sem o retreino das classes originais.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	**89.3**	0.3	7.6	1.6	0.3	0	0	0.9
B	0.2	**79.2**	4.6	2.4	1.2	0.8	4.9	6.7
C	10	3.1	**71.3**	0.3	1.2	0.2	1.0	12.9
D	0.2	2.4	**15.9**	**79.2**	0.2	0.4	0.6	1.2
E	2.0	1.8	**13.9**	0.2	**76.4**	0.2	0.1	**5.4**
F	**11.7**	0.0	**24.6**	0.1	0.8	**48.5**	0.0	**14.3**
G	0.0	1.8	0.9	0.4	0.0	0.1	**95.9**	0.9
H	0.6	1.6	2.7	0.4	0.5	0.4	0.1	**93.7**

vendo, somente, a classe C; uma segunda, com as classes C e H, e por fim, a última, com as classes C, H e A. Os resultados obtidos são apresentados na Tabela 6.10, que também incluiu alguns resultados anteriores visando facilitar as comparações.

Tabela 6.10: Eficiências classe-a-classe e SP, na integração de novas classes, considerando diferentes possibilidades quanto ao retreino dos classificadores originais.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Eficiência (SP)**
Sem retreino	89,3	79,2	71,3	79,2	76,4	48,5	95,9	93,7	$78,5 \pm 0,7$
Retreino C	89,6	80,5	69,0	91,7	88,0	68,7	96,2	94,2	$84,4 \pm 0,5$
Retreino C e H	90,0	77,0	69,2	92,5	91,6	80,4	96,7	92,9	$86,0 \pm 0,5$
Retreino C, H e A	87,0	76,7	70,2	92,7	92,3	92,3	96,7	93,5	$87,5 \pm 0,4$
Retreino completo	87,6	75,9	70,2	94,0	94,0	91,9	96,7	92,2	$87,6 \pm 0,4$

Ao realizar o retreino do especialista C, há uma ganho de 20,2 pontos percentuais na detecção da classe F, e, em torno de 12 pontos, para as classes D e E. Ao considerar, também, o retreino do especialista H, a detecção da classe F eleva-se em 11,7 pontos percentuais, com pequeno ganho para as classes E (3,6 pontos) e D (1,2 pontos). Por fim, caso o especialista da classe A seja também retreinado, há um novo ganho de 11,9 pontos para a classe F.

Resultados anteriores mostram que a técnica de redes classe-especialistas per-

mite a constituição de sistemas classificadores classe-expansíveis robustos, para os quais a inserção de novas classes mostrou-se viável, sem prejuízo expressivo à detecção das classes originalmente consideradas no projeto do sistema. Na inserção de novas classes, com exceção da classe F, foram obtidas eficiências de detecção superiores a $76,4\%$ para todas as classes, mesmo sem o retreino de nenhum especialista, resultado bastante expressivo. Para a classe F, o retreino de parte dos módulos identificados como mais problemáticos, permitiu ganhos expressivos em sua detecção. Acredita-se que caso sejam desenvolvidos mecanismos para prover um melhor encapsulapmento das classes por seus especialistas, resultados ainda melhores possam ser obtidos.

Capítulo 7

Conclusões e Trabalhos Futuros

No contexto de operação dos submarinos, é de especial importância a classificação de contatos. Prover os operadores de sonar com um sistema de apoio à classificação é bastante relevante, visto que isto reduz a carga de estresse e trabalho, e resulta em classificações mais rápidas e precisas, o que acelera o processo de tomada de decisão, vital em cenários de conflito. Um dos objetivos deste trabalho é contemplar alguns dos estágios necessários à produção de um sistema de classificação automática de contatos, na qualidade de um equipamento para operação embarcada, desenvolvido com tecnologia própria, a ser incorporado ao sistema de sonar de submarinos da Marinha Brasileira.

A classificação de contatos é um problema bastante complexo, tendo em vista a multiplicidade de cenários, reflexo de diferentes classes, navios, condições operativas e ambientais possíveis. Normalmente, os sinais envolvidos possuem alta flutuação estatística, dimensionalidade elevada e nível de ruído (próprio e de fundo) considerável. Especialistas apontam, ainda, a necessidade de detectar de 30 a 40 classes de navios.

O desenvolvimento de um sistema de classificação eficiente presume a existência de uma base de dados que caracterize, apropriadamente, todos os cenários operativos. A obtenção desta base, no entanto, é um processo bastante difícil, sujeito a diferentes restrições práticas, tais como: a disponibilidade de navios, o amplo leque de condições operativas e ambientais, o custo das operações e o tempo envolvido. Este trabalho se baseou num banco de dados disponibilizado pelo Instituto de Pesquisas da Marinha (IPqM), constituído por 263 corridas de 28 navios pertencen-

tes à 8 classes distintas adquiridas em raia acústica. Ainda que o conjunto utilizado seja expressivo para o problema em questão, são esperadas restrições estatísticas, em especial, com repeito ao número de classes e às condições operativas contempladas.

Entre requisitos críticos ao sistema de classificação proposto, um dos mais importantes é a capacidade de generalização, visto que a probabilidade do sistema em se deparar com condições distintas às exploradas em seu projeto é alta. Para que seja um útil instrumento de apoio, o sistema deve se mostrar robusto e confiável, apresentando uma alta eficiência, mesmo na identificação de um número tão expressivo de classes distintas. É também relevante a existência de mecanismos para a identificação de classes desconhecidas, e que a inclusão de novas classes possa ser realizada sob o mínimo risco ao bom reconhecimento das classes conhecidas. Por fim, como o sistema operará em tempo real, o custo computacional das técnicas empregadas deve ser considerado.

Em conformidade com diferentes trabalhos anteriores, optou-se por realizar a classificação dos contatos com base nas características espectrais do sinal captado pelo sistema de sonar. Para tal tarefa, utiliza-se a combinação de um sistema de pré-processamento, um sistema extrator de características e um classificador neural.

Em linhas gerais, o papel do pré-processamento consiste em remover do sinal de sonar o ruído de fundo do ambiente de medição, e fornecer janelas espectrais na faixa de freqüência de interesse. Sua escolha foi baseada em resultados de outros trabalhos. Através deste pré-processamento, uma classificação deve ser produzida a cada 0,2 S, tempo compatível com a implementação em tempo real do sistema proposto.

Por outro lado, a extração de características resulta em espaços de dados de menor complexidade, o que simplifica o aprendizado, aspecto útil tendo em vista as restrições estatísticas existentes, e resulta, freqüentemente, em classificadores mais eficientes, com menor custo computacional, logo mais adequados à necessária operação em tempo real do sistema. Extratores de características adaptáveis permitem ainda que o sistema de classificação possa ajustar-se, em operação, a novos cenários operativos, incorporando, por exemplo, novas classes. Quanto ao sistema classificador, classificadores modulares baseados em redes neurais são especialmente indicados, tendo em vista que apresentam uma classificação eficiente, mesmo para

um número expressivo de classes e para eventos de dimensionalidade elevada, além de possuir facilidades quanto à adaptação, em especial, na incorporação de novas classes.

Para a constituição do sistema extrator de características adaptável, optou-se pela técnica de análise de componentes principais, visto sua eficácia comprovada, simplicidade quanto à implementação e operação, e por dispor de algoritmos consagrados de extração em tempo-real. Em trabalhos anteriores, componentes principais se mostraram bem sucedidas no projeto de classificadores compactos para o ambiente de sonar. O número expressivo de técnicas existentes, assim como a carência de análises comparativas baseadas em dados reais com as quais a escolha de um algoritmo mais adequado ao problema de sonar pudesse ser orientada, motivou a elaboração de uma detalhada revisão bibliográfica. Segundo esta revisão, as técnicas foram agrupadas em termos do domínio de aplicação, ou seja, se aplicáveis à extração fora (*off-line*) ou durante a operação do sistema (*on-line*), e com respeito à função objetivo, à arquitetura e à técnica de otimização exploradas em sua derivação. Foi ainda proposta uma arquitetura neural genérica, baseada numa hierarquia de células auto-associativas, para a qual, através da aplicação de técnicas de otimização variadas, foram produzidos diferentes algoritmos, aplicáveis tanto à extração *on-line* quanto *off-line* de componentes. Os métodos propostos foram ainda aplicados à extração de componentes principais do conjunto de dados de sonar, apresentando, do ponto de vista da acuidade das direções das componentes extraídas e do custo computacional envolvido, melhores resultados que métodos consagrados da literatura, tais como o GHA e o PASTd.

Em razão da complexidade do classificador, das restrições estatísticas existentes e das exigências quanto à eficácia e à confiabilidade, cuidado especial foi dedicado ao projeto do classificador e a previsão do seu desempenho. Dois cenários foram considerados: a seleção de dados por espectros e a por corridas, a primeira, onde o projeto e a avaliação consideraram um mesmo conjunto de condições operativas; e a segunda, mais realística, onde a avaliação utilizou condições operativas distintas daquelas exploradas no treinamento do classificador.

Para os dois cenários foi realizada uma identificação das classes com maiores restrições estatísticas, a qual pode orientar novas aquisições de dados. Esta

identificação foi baseada no desempenho de generalização de um classificador MLP totalmente conectado, através do qual foram identificadas as classes de menor desempenho e maiores incertezas quanto à classificação, sujeitas, portanto, à maiores restrições quanto sua caracterização estatística. As incertezas de classificação estimadas visaram, ainda, prover uma medida de confiabilidade das classificações providas pelo sistema automático ao operador de sonar. A avaliação da capacidade de generalização utilizou a técnica de subamostragem aleatória, que tende a ser pessimista, ou seja, fornecer estimativas inferiores ao valor real, que é um aspecto atrativo a aplicações militares, e considerou uma partição meio-a-meio dos eventos disponíveis entre os conjuntos de projeto e avaliação, de forma que uma ênfase especial fosse dedicada à avaliação da generalização. Conforme esperado, a avaliação de classificadores produzidos utilizando conjuntos de projeto e avaliação baseados na seleção por corridas resultou em menores valores de eficiências e maiores flutuações de desempenho. Para a seleção por espectros, foram obtidas eficiências SP de $(92,1 \pm 0,5)\%$ e $(95,6 \pm 0,2)\%$, enquanto na seleção por corridas de $(80,6 \pm 1,6)\%$ e $(83,1 \pm 0,9)\%$, para redes com 10 e 40 neurônios na camada intermediária, respectivamente. As classes A, B, C e H, em especial, a classe C, para ambas modalidades de seleção, foram identificadas como as mais críticas.

Outro requisito relevante ao desenvolvimento do sistema de classificação é prover conjuntos de projeto e avaliação que reflitam, apropriadamente, os diferentes cenários de operação existentes, sob risco de comprometer a produção do classificador ou resultar em estimativas não realísticas de seu desempenho. Uma das propostas deste trabalho é utilizar agrupamentos para esta seleção, visto que são estruturas que particionam os dados em grupos de características estatísticas similares, o que permite uma melhor constituição de ambos conjuntos. Em razão do número expressivo de classes e eventos disponíveis, optou-se, por questões de simplicidade, custo computacional e eficácia, pelas utilização das técnicas de agrupamento seqüencial e hierárquico, produzindo-se um agrupamento por classe. Uma vez que as técnicas de agrupamento costumam apresentar melhor desempenho em espaços de dimensão reduzida, e como os eventos possuem dimensionalidade elevada, foi considerada a produção dos agrupamentos com base em eventos compactados através das direções fornecidas pelas análises de componentes principais (PCA) e de discriminação

(PCD). Como o valor do raio vigilância, para o agrupamento seqüencial, e o nível de corte, para o agrupamento hierárquico, possuem relação direta com o número e as características estatísticas dos grupos por eles identificados, foram propostas ferramentas e critérios para auxiliar a definição destes parâmetros, considerando tanto a seleção por espectros quanto por corridas.

Em relação à seleção por espectros, para o agrupamento seqüencial, foi proposto um critério onde, tendo como base uma curva que relaciona o número de grupos do agrupamento como função do valor do raio de vigilância, definido por uma fração da moda da distância euclidiana dos eventos da classe, são identificados um conjunto de candidatos ao raio. Para este conjunto, o valor de raio mais adequado é identificado utilizando o próprio classificador como figura de mérito. Ao considerar frações da moda para a definição dos raios, uma vantagem é que o seu processo de seleção para as diferentes classes é uniformizado. Foi considerada, também, a utilização de critérios relativos, processo onde índices estatísticos identificam, dentre um conjunto de agrupamentos produzidos para diferentes escolhas de raio, qual deles melhor reflete a estrutura existente nos dados. Para o agrupamento hierárquico, é proposto um critério onde, através da modelagem matemática das curvas de dissimilaridade de formação do agrupamento, realiza-se um corte baseado na idéia de constantes de decaimento. Este procedimento é interessante, pois explora a estrutura existente nos dados para a definição do corte, o qual é uniformizado, também, em relação às diferentes classes. Por este critério, de forma similar ao anterior, o corte mais apropriado é identificado dentre um conjunto de candidatos utilizando o classificador como figura de mérito. Para a seleção destes candidatos, é proposta a elaboração e inspeção de uma curva que relaciona o número de grupos definidos pelo corte segundo diferentes constantes de tempo.

Resultados referentes à aplicação de agrupamentos para a seleção por espectros mostraram que, para a técnica seqüencial, maiores eficiências de generalização foram obtidas para os classificadores cujos conjuntos de projeto e avaliação foram selecionados através de agrupamentos produzidos sobre dados compactados em 10 componentes de discriminação, resultado coerente, visto que esta técnica privilegia informação discriminante dos dados. Com respeito à seleção dos raios, a análise considerou, inicialmente, três granularidades subjetivas: fina, intermediária e gros-

seira, correspondentes a escolhas de raio de 0,5, 1,0 e 5,0 vezes a moda. O melhor resultado foi obtido ao considerar uma granularidade fina para todas as classes. Posteriormente, foi avaliada a definição dos valores de raio classe-a-classe através de 3 índices estatísticos: *Silhouette*, *Dunn* e *Davies-e-Bouldin*, que consideraram 12 valores de raio escolhidos entre 0,5 a 5,0 vezes a moda. Para um total de 5 classes, um valor de raio reduzido, correspondente a 0,5 vezes a moda, foi identificado como o mais apropriado. Para as demais, tem-se valores de raio médio (3,0) e grande (5,0). No agrupamento hierárquico, três constantes de decaimento (0,5, 0,9 e 1,2) foram consideradas, as quais resultaram em agrupamentos subjetivamente classificados como de granularidades grosseira, intermediária e fina. Em oposição ao agrupamento seqüencial, um melhor desempenho foi verificado para a granularidade baixa. Na seleção dos conjuntos, o agrupamento seqüencial produziu resultados superiores ao hierárquico. Quando comparadas a seleção dos raios do agrupamento seqüencial por granularidade subjetiva ou através de índices, verifica-se um melhor desempenho da primeira, que produziu classificadores com eficiência SP de $(92,6 \pm 0,3)\%$ e $(96,2 \pm 0,2)\%$, considerando 10 e 40 neurônios na camada intermediária do classificador, respectivamente.

Na seleção por corridas, como a formação dos conjuntos de projeto e avaliação deve considerar grupos de corridas com características estatísticas similares, é proposto um critério onde cada corrida é associada a um vetor representativo, sendo produzido um agrupamento com base nos diferentes vetores gerados. Corridas cujos vetores, neste agrupamento, pertençam a um mesmo grupo, são tidas como similares. A produção de vetores representativos é baseada no agrupamento gerado para a seleção por espectros, e considerou 2 critérios que exploram as características relativas à forma como as corridas estão distribuídas pelos grupos deste agrupamento: uma primeira, onde o vetor representativo é definido pelo baricentro dos centros dos grupos que contém eventos da corrida; e uma segunda, onde se define um vetor binário, referido como vetor de pertinência, cujo valor de cada componente indica se a corrida possui ou não eventos no interior do grupo a ela correspondente. Para a geração do agrupamento de vetores representativos é utilizada a técnica hierárquica, com seleção do nível de corte realizado pelo critério de constantes de decaimento anteriormente proposto. O melhor resultado é obtido pelo agrupamento dos vetores de

pertinência, o qual resultou em eficiências SP de $(84,3\pm0,5)\%$ e $(87,5\pm0,4)\%$, para classificadores com 10 e 40 neurônios na camada intermediária, respectivamente.

Estratégia útil ao desenvolvimento do sistema classificador é dividir-e-conquistar, ou seja, subdividir a classificação em tarefas mais simples, cuja solução pode explorar diferentes técnicas, escolhidas de acordo com as especificidades de cada tarefa. Utilizar múltiplos classificadores pode ainda resultar num sistema mais seguro e robusto, para o qual a inclusão de novos cenários é viável. Este trabalho propõe a constituição do sistema classificador por redes modulares classe-especialistas. Nesta proposta, um classificador MLP de 2 camadas, totalmente conectado, com saída binária, é associado a cada classe, possuindo o papel de identificar se o evento pertence ou não a classe a ele correspondente. Entre atrativos desta técnica, tem-se um classificador eficiente, mesmo em problemas com alta-dimensionalidade, escalável e com capacidade de adaptação parcial, o que é útil na incorporação de novos cenários operativos. Pela modularidade, a detecção e a inclusão de novas classes pode ainda ser facilmente realizada.

Para esta proposta são discutidos o treinamento, a escolha da topologia das redes classe-especialistas e o critério de integração explorados na formação do sistema classificador. Em relação ao treinamento, ênfase especial é dedicada a função objetivo e a seleção da parada. Para a função objetivo são apresentadas 3 alternativas: batelada uniforme, classe-e-não-classe e classe-e-não-classe normalizada; para a seleção do ponto de parada do treinamento e da topologia, tem-se 5 propostas de mérito: eficiências de detecção, falso-alarme e SP, área da ROC e valor do erro quadrático médio. Com respeito à combinação das redes classe-especialistas na formação do sistema classificador, 4 critérios são avaliados: por máximo, que define a classe pelo especialista de maior valor de saída; linear um e dois, onde a classe é definida com base numa combinação linear das saídas dos especialistas, e, por fim, não-linear, onde a decisão é realizada por uma rede neural que utiliza as características extraídas por cada especialista. É discutida, também, a identificação e inserção de novas classes. Para a primeira, são propostos critérios baseados na comparação das saídas com 1 ou 2 limiares, ou na produção de agrupamentos baseados em seus valores. Para o último, é proposto definir um grupo, por hiperesfera ou hiperelípsóide, para cada classe, identificando os eventos como pertencentes à dada

classe apenas se estiverem situados no interior do grupo a ela correspondente. Em relação à inserção de classes, o procedimento consiste em inserir novos especialistas ao sistema, realizando, se necessário, o retreino dos especialistas originais e o ajuste do critério de integração.

As análises desenvolvidas para os especialistas utilizaram conjuntos de projeto e avaliação identificados para a seleção por corridas, segundo a técnica de agrupamento de vetores de excitação. Em relação à função objetivo, o melhor resultado foi obtido pela função classe-e-não-classe normalizada. Quanto ao critério de parada, tem-se a eficiência SP. Estes resultados sinalizam que, para um melhor desempenho do sistema final, os classificadores classe-especialistas devem conjugar habilidades tanto na detecção de eventos de suas classes quanto na rejeição de eventos das demais. Para escolha da topologia, o índice de melhor desempenho foi a área da ROC, resultando em classificadores especialistas com 3 (classes A e G) a 78 neurônios (classe C) na camada intermediária, eficiência mínima de $70,2\%$ (classe C) e máxima de 94% (classes D e E), com desempenho global de $(87,6 \pm 0,4)\%$ quanto à eficiência SP. Em relação a integração, o melhor resultado foi obtido pela integração não-linear, para a qual foi utilizada uma rede neural com 184 nós de entrada, 26 neurônios na camada escondida e 8 neurônios na camada de saída, alimentada pelas saídas das camadas intermediárias das especialistas. Esta arquitetura apresentou uma eficiência SP de $(88,5 \pm 0,4)\%$. Em segundo lugar, tem-se o critério de máximo, com eficiência SP de $(87,6 \pm 0,4)\%$.

Para avaliar a capacidade de detecção e inclusão de novas classes ao sistema classificador baseado em redes classe-especialistas, o sistema foi projetado utilizando apenas as classes A, B, C e H, uma vez que este conjunto foi identificado, anteriormente, como de classificação mais crítica. Nesta análise, por maior simplicidade, o critério de máximo foi explorado para a combinação dos especialistas. Entre as propostas de detecção de novas classes, o critério que realiza a comparação do maior valor de saída dos especialistas com um limiar arbitrário apresentou o melhor desempenho, permitindo uma eficiência de detecção mímima de $66,8\%$ (classe B) e máxima de $88,5\%$ (classe H), identificando, no mínimo, para a classe F, 21%, e, no máximo, para a classe G, 61,2% dos eventos como provenientes de novas classes, caso um limiar de valor 0,95 fosse escolhido. Para os eventos de novas classes não

identificados, parcela significativa é classificada como proveniente da classe C, com destaque às classes D, E e F; da classe A (eventos da classe F) e da classe B (eventos da classe G). Quanto à inclusão das novas classes ao sistema, é possível realizá-la considerando ou não o retreino de parte ou todos classificadores originalmente existentes no sistema. Para uma inclusão sem retreino de nenhum especialista original, o sistema apresentou eficiências superiores a 70% para todas as classes, exceto para a classe F, com eficiência de $48,5\%$. Foi mostrado que ao realizar o retreinamento da classe C, a eficiência da classe F sobe para $68,7\%$, e o sistema apresenta eficiências superiores a 88% para as demais classes. Considerando, também, o retreino das classes A e H, é possível elevar a detecção da classe F ao patamar de $91,9\%$. Resultados demonstraram que, caso o encapsulamento das classes por seus especialistas seja melhorado, um melhor desempenho quanto à detecção e a inclusão de novas classes possa ser obtido.

Os resultados obtidos neste trabalho demonstram que os agrupamentos constituem um eficiente instrumento para a seleção de conjuntos de projeto e avaliação em sistemas de classificação com severas restrições quanto à capacidade de generalização ou na caracterização estatística das classes envolvidas. Neste caso, a utilização de redes classe-especialistas mostrou-se especialmente indicada, tendo em vista os expressivos valores de eficiência obtidos para as diferentes classes, assim como os promissores resultados na identificação e incorporação de novas classes ao sistema.

7.1 Trabalhos futuros

O bom desempenho obtido para a técnica de classificadores classe-especialistas sinaliza que a estratégia de dividir-e-conquistar é atrativa ao problema de sonar, conforme esperado. Assim, a classificação de classes mais críticas, tais como a classe C, pode explorar as máquinas de comitê, técnica onde vários classificadores são dedicados à solução de uma mesma tarefa. Neste caso, faz-se necessário um estudo detalhado quanto à produção e a integração dos diferentes classificadores do comitê, assim como uma discussão de quais classes serão envolvidas e como este comitê será incorporado ao sistema especialista.

Caso o encapsulamento das classes seja melhorado, maiores eficiências quanto

à identificação e a incorporação de novas classes podem ser obtidas. Uma possibilidade é a pesquisa de mecanismos para um melhor encapsulamento das classes pelas redes classe-especialistas que sejam aplicáveis a dados de dimensionalidade elevada. Pode-se, ainda, investigar técnicas alternativas para a constituição destes classificadores. Uma possibilidade é utilizar um enfoque híbrido, onde agrupamentos e classificadores neurais sejam utilizados para esta tarefa. Como ponto de partida desta proposta pode-se explorar os agrupamentos produzidos para a seleção dos conjuntos de projeto e avaliação do classificador. Outra alternativa interessante são os classificadores baseados em curvas principais, para os quais podem ser explorados critérios mais sofisticados de classficação que a distância à curva representativa de cada classe.

De posse dos agrupamentos produzidos para as seleções por espectros e corridas, pode-se ainda realizar uma análise prospectiva dos dados, visando identificar as características dos grupos formados. Nesta análise, por exemplo, poderiam ser identificadas quais condições operativas caracterizam cada grupo, o que permitiria identificar, para corridas distintas, quais condições de maquinário são comuns a ambas. Este estudo poderia basear uma pesquisa de mecanismos para a caracterização das condições operativas dos contatos, resultando num sistema de apoio tático que, além de fornecer ao operador a classe a qual a embarcação pertence, proveria informações de interesse militar tais como o tipo de propulsão, o número de eixos e pás e a velocidade do contato.

O conjunto de dados explorado neste trabalho considerou aquisições de um número restrito de classes e cenários operativos, realizada em raia acústica, que possui baixa profundidade, por meio de um hidrofone junto ao fundo da raia. Em cenários reais de operação, o sinal seria adquirido em águas profundas, por meio de um conjunto de hidrofones, cuja composição, referida como formação do feixe, resultaria no sinal de entrada do sistema classificador. Quanto ao sinal recebido, características estatísticas diferentes são esperadas, tanto pelas condições de propagação distintas do meio quanto pela contaminação do sinal do contato pelo ruído do maquinário em operação no submarino, referido como ruído próprio. Há ainda a possibilidade do sistema em se deparar com mais de um contato simultaneamente, com novos navios de classes conhecidas, ou mesmo, com embarcações de novas clas-

ses. Assim, uma importante etapa futura consiste em avaliar o sistema proposto em cenários reais de operação, para os quais é importante discutir mecanismos para a detecção dos sinais envolvidos, visto que este trabalho e anteriores concentraram-se apenas na sua classificação. Pesquisas tem sido desenvolvidas visando à separação dos ruídos próprio e do contato, assim como para isolar um contato de interesse na ocorrência de múltiplos contatos simultâneos. Resultados promissores estão sendo obtidos pela utilização de componentes fornecidas pela análise de componentes independentes [66], tecnica capaz de separar um ou mais sinais oriundos de fontes estatísticas independentes que tenham sido misturados. Esta técnica poderia também ser explorada visando à decomposição dos sinais dos contatos, o que seria útil à caracterização do maquinário em operação no seu interior, podendo também constituir importante instrumento de auxilio à classificação.

Sistemas de classificação que possuam a capacidade de adaptar-se, em operação, à novos cenários são de especial interesse para o problema de sonar passivo. Um primeiro passo para o desenvolvimento deste sistema foi discutido neste trabalho, ao realizar-se um estudo de sistemas extratores de características adaptativas, focando-se na extração de componentes principais de representação (PCA). Este é um problema bastante complexo, o qual demanda extensas pesquisas quanto à identificação eficiente de novas classes, e com respeito ao processo de atualização do sistema classificador, o qual pode ser semi-automático, ou seja, com interferência do operador, ou totalmente automático.

Apêndice A

Artigos Publicados

Nesta seção, apresentamos os artigos publicados em congressos nacionais e internacionais relacionados ao desenvolvimento desta tese. Cada artigo é acompanhado de breve descrição.

1. SOUZA FILHO, J. B. O., SEIXAS, J. M., "IR-PCA: Um Algoritmo Neural Acurado e Rápido para a Extração de Componentes Principais". In: *XVI CBA - Congresso Brasileiro de Automática*, Salvador, Bahia, Brasil, 2006.
   Neste trabalho, um novo método de extração *off-line* de componentes principais por redes neurais é proposto. Aliando uma arquitetura escalável, de fácil implementação, o método possui baixo custo computacional por iteração e apresenta, em termos de acurácia e custo computacional, um desempenho significativamente superior a algortimos consagrados da literatura.

2. SOUZA FILHO, J. B. O., SEIXAS, J. M., "Seleção Estatística de Dados para Classificadores Neurais de Sinais de Sonar Passivo". In: *VII CBRN - Congresso Brasileiro de Redes Neurais*, Natal, RN, Brasil, 2005.
   Neste trabalho discute-se a seleção de dados para o projeto de classificadores em problemas sujeitos a restrições estatísticas do banco de dados. Considerando a classificação de espectros de sonar passivo, provenientes de 25 navios pertencentes a 8 classes distintas, é mostrado que a seleção dos conjuntos de projeto e teste por meio de técnicas de análise de agrupamentos permite, com um menor custo computacional, obter classificadores de maior eficiência média e melhor desempenho para as classes mais críticas. Utilizando esta técnica, um

classificador com uma eficiência média de $94,6\%$ foi obtido, o que é significativo para o problema em questão.

3. FERNANDES, H. L., SEIXAS, J. M., NEVES, S. R., SOUZA FILHO, J. B. O.,"Combining Morphological Mapping and Principal Curves for Ship Classification". In: *IEEE International Symposium on Signals, Circuits and Systems (ISSCS)* , Romênia, 2005.

   É desenvolvido um classificador de navios que utiliza curvas principais para extrair informação relevante de imagens segmentadas. O classificador é baseado na distância euclidiana do ponto cujas coordenadas representam as características relevantes extraídas da imagem de entrada às curvas atribuídas à cada classe. Esta metodologia é atrativa, dado que possui baixo custo computacional na fase de extração, sendo facilmente escalável para um número arbitrário de classes. Foi antigida uma eficiência média de classificação de $97,3\%$, o que supera resultados prévios baseados em redes neurais.

4. SOUZA FILHO, J. B. O., SEIXAS, J. M., "Agrupamento Estatístico de Dados de Sonar Passivo para o Projeto de Classificadores". In: *I EPAS - Encontro de Propagação em Acústica Submarina*, Rio de Janeiro, RJ, Brasil, 2004.

   Este trabalho explora as técnicas de validação cruzada, redes ART e clusterização hierárquica para a seleção estatística de espectros relevantes do conjunto de dados. Critérios para o dimensionamento dos agrupamentos são propostos. Utilizando um classificador MLP baseado na técnica de componentes principais de discriminação, os diferentes métodos de agrupamento são comparados com base na eficiência de generalização.

5. FERNANDES, H. L., SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificação de Sinais Acústicos Submarinos Utilizando Curvas Principais". In: *XV CBA - Congresso Brasileiro de Automática*, Gramado, RS, Brasil, 2004.

   Este artigo propõe a classificação de contatos utilizando curvas principais para a caracterização das diferentes classes de navios. É analisado o impacto da complexidade da curva e da metodologia de normalização dos dados na eficiência de generalização do classificador.

6. FERNANDES, H. L., SOUZA FILHO, J. B. O., SEIXAS, J. M., "Extração de Características de Sinais de Sonar Passivo Usando Curvas Principais". In: *I EPAS - Encontro de Propagação Acústica Submarina*, Rio de Janeiro, RJ, Brasil, 2004.

   Neste trabalho, avaliam-se as curvas principais como metodologia de extração de características para os sinais de sonar. É realizada uma análise da distância entre as curvas, segundo diferentes métricas e metodologias, visando identificar as classes de classificação mais crítica.

7. SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificação de Sinais de Sonar Passivo baseada em Filtragem Casada". In: *XV CBA - Congresso Brasileiro de Automática*, Gramado, RS, Brasil, 2004.

   Este trabalho propõe um classificador de contatos pela técnica de filtragem casada. É apresentada uma nova modelagem para o problema, na qual não se faz necessária a caracterização estatística do ruído. A otimização do classificador, seja pela compactação dos sinais ou pela seleção de dimensões discriminantes, é também discutida.

8. COSTA, B. F. P., SOUZA FILHO, J. B. O., SEIXAS, J. M., "Passive Sonar Signal Classification Using Expert Neural Networks". In: *SBRN 2004 - Braziliam Symposium on Artificial Neural Networks*, São Luiz, MA, Brasil, 2004.

   Propõe-se a classificação de contatos por redes especialistas. Segundo esta metodologia, utilizando a estratégia de dividir-e-conquistar, forma-se um classificador pelo agrupamento de classificadores mais simples, especializados em cada classe. Os resultados mostraram que este enfoque é interessante, permitindo ao classificador absorver novas classes.

9. COSTA, B. F. P., SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificação Sonar Passiva Utilizando Redes Neurais Especialistas". In: *I EPAS - Encontro de Propagação Acústica Submarina*, Rio de Janeiro, RJ, Brasil, 2004.

   Utilizando a metodologia anteriormente descrita, aspectos relativos à metodologia de treinamento e seu impacto na generalização, assim como o dimensionamento das redes especialistas são discutidos.

10. SOUZA FILHO, J. B. O., CALôBA, L. P., SEIXAS, J. M., "An Accurate and Fast Neural Method for PCA Extraction". In: *IJCNN2003 - International Joint Conference on Neural Networks*, Portland, Oregon, USA, 2003.
    Neste artigo é proposto um novo método *on-line* para a extração de componentes principais. Testes mostram que o método possui maior acurácia e menor tempo de convergência que métodos consagrados da literatura.

11. SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificador de Contatos baseado em Filtragem Casada". In: *II Workshop em Acústica Submarina*, Rio de Janeiro, RJ, Brasil, 2002.
    Propõe uma nova modelagem ao problema e apresenta os primeiros resultados relativos à tecnica desenvolvida, em maiores detalhes, na referência: *Classificação de Sinais de Sonar Passivo baseada em Filtragem Casada*, descrita acima.

12. SOUZA FILHO, J. B. O., SEIXAS, J. M., "Curvas Principais na Classificação de Sinais de Sonar Passivo". In: *II Workshop em Acústica Submarina*, Rio de Janeiro, RJ, Brasil, 2002.
    Propõe-se um classificador de contatos baseado na modelagem de cada classe por uma curva principal. Este classificador é baseado na distância de cada evento à curva representativa de cada classe. A despeito da simplicidade do critério de classificação, bons resultados são obtidos.

13. SOUZA FILHO, J. B. O., SOUZA, M., CALÔBA, L. P., SEIXAS, J. M., "Extração Neural de Componentes Principais em Aplicações de Elevada Dimensionalidade". In: *V SBAI - Simpósio Brasileiro de Automação Inteligente*, Canela, RS, Brasil, 2001.
    Neste trabalho é proposto um novo método de extração de componentes principais, que é baseado numa rede MLP auto-associativa submetida a um treinamento diferenciado. Testes de acuidade e custo computacional mostram um melhor desempenho do método em relação a algoritmos consagrados da literatura.

14. SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificador Neural Online para Sonar Passivo utilizando um Processador Digital de Sinais de Alto-Desempenho".

In: *V CBRN - Congresso Brasileiro de Redes Neurais*, Rio de Janeiro, RJ, Brasil, 2001.

Neste trabalho é discutida a implementação de um sistema de classificação neural *on-line* para sinais de sonar passivo utilizando um processador digital de sinais de alto-desempenho, o ADSP21062. O sistema foi codificado em linguagem Assembly, otimizado quanto a velocidade de execução, sendo acessível por uma interface amigável através de um microcomputador padrão IBM-PC.

Apêndice B

Técnicas de redução dimensional

Para número expressivo de aplicações, é necessário lidar com conjuntos de dados de dimensão elevada, logo a seleção das variáveis relevantes à solução do problema é relevante, senão fundamental. Para aplicações que operam em tempo real, tal como o sonar passivo, a questão da dimensionalidade é ainda mais crítica, dadas as restrições quanto ao volume de processamento. Conclui-se que a extração de características relevantes pode ser fundamental para a viabilidade de uma aplicação.

Dado um evento n-dimensional $\mathbf{x}$, uma característica arbitrária (F_i) de $\mathbf{x}$ pode ser expressa na forma:

$$F_i = F(\mathbf{x}) = F(x_1, \ldots, x_n), \tag{B.1}$$

onde $F(\mathbf{x})$ representa uma função de extração de características, também arbitrária, onde $x_1, \ldots, x_n$ correspondem às componentes do vetor $\mathbf{x}$. Um caso particular da função F_i é uma função de extração de características lineares, para a qual F_i assume a forma:

$$F_i = \alpha_{i1}x_1 + \ldots + \alpha_{in}x_n, \tag{B.2}$$

onde $\alpha_{i1} \ldots \alpha_{in}$ são coeficientes a serem obtidos por técnica apropriada. Uma outra forma de representar a Equação B.2 é pelo produto escalar $F_i = \mathbf{a_i}^T\mathbf{x}$, sendo $\mathbf{a}_i$ um vetor coluna de componentes dadas pelos coeficientes α_i.

Como, normalmente, mais de uma característica é utilizada, pode-se escrever que o conjunto das características é definido por um vetor coluna $\mathbf{f}$, de dimensão k, dado por:

$$\mathbf{f} = \begin{bmatrix} F_1 & \ldots & F_k \end{bmatrix}^T = \begin{bmatrix} \mathbf{a}_1 & \ldots & \mathbf{a}_k \end{bmatrix}^T \mathbf{x} = \mathbf{T}\mathbf{x}, \tag{B.3}$$

onde $\mathbf{T}$ corresponde a matriz de extração de características, ou mesmo, de compactação dos dados, pois reduz a dimensão dos eventos de n para k, onde k corresponde ao número de características selecionadas.

Propor uma matriz $\mathbf{T}$ é um processo dependente do tipo de informação e da metodologia utilizada para sua determinação. A seguir, discutiremos algumas técnicas para sua obtenção.

B.1 Análise de componentes principais

A análise de componentes principais - do inglês: *Principal Component Analysis* - PCA - [4, 67, 157] é uma técnica de análise multivariável consagrada, extensivamente aplicada em variadas áreas, entre elas: detecção, estimação, reconhecimento de padrões, processamento de áudio e imagens, assim como na compactação de espaços de elevada dimensionalidade. Esta técnica é baseada na expansão de processos aleatórios em série de *Karhunen-Loève* (KL) [157].

Considere $\mathbf{x}$ um vetor aleatório de dimensão N, formado pela amostragem periódica de um processo estocástico contínuo, de média nula, realizada durante um período de tempo arbitrário T. O objetivo da PCA discreta é obter os coeficientes c_i e os vetores base $\boldsymbol{\phi_i}$ de forma que $\mathbf{x}$ possa ser escrito na forma [27]:

$$\mathbf{x} = \sum_{i=1}^{N} c_i \boldsymbol{\phi_i}, \tag{B.4}$$

onde c_i são coeficientes aleatórios, dependentes da realização em questão, e $\boldsymbol{\phi_i}$ constituem os vetores base determinísticos, os quais estão associados à expansão do processo gerador da realização.

Para a determinação de c_i e $\boldsymbol{\phi_i}$ serão considerados os seguintes critérios:

1. Média quadrática como critério de convergência:

$$EQ = E\left[\left(\mathbf{x} - \sum_{i=1}^{N} c_i \boldsymbol{\phi_i}\right)^2\right] = 0, \tag{B.5}$$

   isto é, os coeficientes e vetores base consistem nos zeros da função EQ, ou ainda, o erro quadrático cometido na expansão do processo através de N funções base é nulo.

2. Coeficientes da expansão determinados por projeção:

$$c_i = \boldsymbol{\phi}_{\boldsymbol{i}}^T \mathbf{x} \tag{B.6}$$

3. Funções base ortogonais e com energia unitária [1]:

$$\boldsymbol{\phi}_{\boldsymbol{i}}^T \boldsymbol{\phi}_{\boldsymbol{j}} = \delta_{ij} \tag{B.7}$$

4. Coeficientes da expansão descorrelacionados:

$$E[c_i c_j] = \lambda_i \delta_{ij}, \tag{B.8}$$

onde λ_i é uma constante arbitrária.

Para as condições acima relacionadas, pode-se mostrar que os vetores $\boldsymbol{\phi_i}$ devem satisfazer [157]:

$$\mathbf{R_x} \boldsymbol{\phi_i} = \lambda_i \boldsymbol{\phi_i}, \tag{B.9}$$

ou seja, os vetores base da expansão correspondem aos N autovetores [70] da matriz de correlação $\mathbf{R_x}$, a qual é dada por:

$$\mathbf{R_x} = E[\mathbf{x}\mathbf{x}^T] \tag{B.10}$$

Da Equação B.4 pode-se verificar que o processo aleatório $\mathbf{x}$ pode ser expandido, de forma aproximada, utilizando um subconjunto S dos N vetores base disponíveis, isto é:

$$\tilde{\mathbf{x}} = \sum_{i \in S} c_i \boldsymbol{\phi_i} \tag{B.11}$$

O erro quadrático médio da expansão de $\tilde{\mathbf{x}}$ pode ser escrito na forma:

$$\begin{aligned} EMQ =& E\left[\left(\mathbf{x} - \sum_{i \in S} c_i \boldsymbol{\phi_i}\right)^2\right] = \left[\left(\sum_{i=1}^{N} c_i \boldsymbol{\phi_i} - \sum_{j \in S} c_j \boldsymbol{\phi_j}\right)^2\right] \\ EMQ =& E\left[\left(\sum_{i \notin S} c_i \boldsymbol{\phi_i}\right)^2\right] = E\left[\sum_{i \notin S} \sum_{j \notin S} c_i c_j \boldsymbol{\phi}_{\boldsymbol{i}}^T \boldsymbol{\phi_j}\right] \\ EMQ =& \sum_{i \notin S} E[c_i^2] = \sum_{i \notin S} \lambda_i, \end{aligned} \tag{B.12}$$

[1] $\delta_{ik} = 1$ para $i = k$ e $\delta_{ik} = 0$ para $i \neq k$.

isto é, o erro cometido na aproximação está relacionado ao somatório dos autovalores associados aos autovetores base não considerados na expansão. Assim, se o conjunto S for formado pelos k autovetores associados aos maiores autovalores, o erro quadrático médio desta expansão, para um valor de k fixo, será mínimo.

Por outro lado, a energia associada ao processo $\tilde{\mathbf{x}}$ é dada por:

$$\begin{aligned} E[|\tilde{\mathbf{x}}|^2] &= E[\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}] = E\left[\sum_{i\in S} c_i\boldsymbol{\phi_i}^T \sum_{j\in S} c_j\boldsymbol{\phi_j}\right] \\ E[|\tilde{\mathbf{x}}|^2] &= \sum_{i\in S} E[c_i^2] = \sum_{i\in S}\lambda_i, \end{aligned} \tag{B.13}$$

donde concluímos que os autovetores, obtidos pela minimização do erro quadrático médio, são aqueles que maximizam o valor quadrático da projeção dos dados em suas direções, ou ainda, que retêm a maior parte da energia do processo original. Como é suposto que o processo possui média nula, estas direções correspondem ainda às direções que maximizam a variância da projeção dos dados.

Definem-se como as componentes principais do processo $\mathbf{x}$, as direções que contribuem de forma mais significativa para a representação do processo $\tilde{\mathbf{x}}$, conforme a Equação B.4. Como foi mostrado que a contribuição de cada autovetor está relacionada a seu autovalor associado, as componentes principais correspondem aos autovetores de $\mathbf{R_x}$ associados aos maiores autovalores, isto é, a primeira componente principal corresponde ao autovetor associado ao maior autovalor; a segunda, ao segundo maior autovalor, e assim, sucessivamente. Como, para variados problemas, a energia do processo se deposita de forma mais concentrada em algumas componentes, restando às demais valores pouco relevantes de energia, ou mesmo, ruído, a PCA fornece um método adequado de compactação linear. Sejam $\mathbf{e}_1, \ldots, \mathbf{e}_k$ os autovetores associados aos k maiores autovalores ordenados de $\mathbf{R_x}$. Retormando a Equação B.3, para a compactação por PCA, a matriz $\mathbf{T}$ é dada por:

$$\mathbf{T} = \begin{bmatrix} \mathbf{e}_1 & \ldots & \mathbf{e}_k \end{bmatrix}^T \tag{B.14}$$

Aplicando a matriz $\mathbf{T}$ proposta pela Equação B.14 na B.3, temos que o vetor

de características $\mathbf{f}$ possui componentes ainda descorrelatadas, pois [2]:

$$\begin{aligned}
\mathbf{R_f} &= \begin{bmatrix}\mathbf{e}_1 & \dots & \mathbf{e}_K\end{bmatrix}^T \mathbf{R_x} \begin{bmatrix}\mathbf{e}_1 & \dots & \mathbf{e}_K\end{bmatrix} \\
\mathbf{R_f} &= \begin{bmatrix}\mathbf{e}_1 & \dots & \mathbf{e}_K\end{bmatrix}^T \begin{bmatrix}\mathbf{e}_1 & \dots & \mathbf{e}_N\end{bmatrix} \mathbf{\Lambda_N} \begin{bmatrix}\mathbf{e}_1 & \dots & \mathbf{e}_N\end{bmatrix}^T \begin{bmatrix}\mathbf{e}_1 & \dots & \mathbf{e}_K\end{bmatrix} \\
\mathbf{R_f} &= \mathbf{\Lambda_K},
\end{aligned} \tag{B.15}$$

onde $\mathbf{\Lambda_k}$ é uma matriz diagonal cujas entradas são dadas por $\lambda_1, \dots, \lambda_k$, ou seja, a análise PCA aplicada a compactação produz variáveis aleatórias descorrelacionadas, as quais são formadas por um pequeno número de combinações lineares das variáveis originais, retendo, no sentido da energia ou, equivalentemente, do erro médio quadrático, o máximo de informação possível do conjunto original [67].

B.1.1 Técnicas para a extração de componentes principais

A extração PCA está relacionada a fatoração da matriz de covariância em autovetores e autovalores. Esta fatoração, usualmente, é realizada por meio de algoritmos clássicos da algebra linear numérica, que promovem a diagonalização de matrizes arbitrárias através da aplicação de sucessivas transformações lineares. Entre estes métodos, os mais comumente aplicados para a diagonalização de matrizes hermitianas são o método de *Jacobi*, a fatoração de *Household*, o método de *Givens* e a fatoração *QR* [88]. Uma descrição mais detalhada destes algoritmos pode ser encontrada em [97].

Ainda que na maior parte das aplicações, apenas algumas componentes sejam necessárias, os métodos clássicos, na sua forma usual, extraem todas as componentes do processo de interesse. Existem, ainda, alguns métodos para a extração de pequeno número de componentes [63]. Um sério inconveniente de todos estes métodos é a necessidade da estimação da matriz de covariância, a qual é contra-indicada ou até mesmo impossível, por restrições de memória ou de volume de processamento, em aplicações de dimensão elevada. Uma alternativa a este problema são os métodos adaptativos que foram discutidos no Capítulo 3.

[2]Foi explorada a expansão de $\mathbf{R_x}$ em termos de: $\mathbf{R_x} = \mathbf{Q\Lambda Q}^T$, onde Λ é uma matriz diagonal formada pelos autovalores de $\mathbf{R_x}$ ordenados de forma decrescente e a matriz $\mathbf{Q}$ possui, como colunas, os autovetores associados a estes autovalores [70].

B.2 Análise de componentes principais de discriminação

Conforme discussão anterior, a análise PCA obtém direções relacionadas a uma representação ótima dos dados. Para fins de classificação, no entanto, a informação mais relevante é aquela que define as diferenças entre as classes. A análise que fornece direções privilegiadas para fins de classificação é a análise de componentes principais de discriminação (*PCD - Principal Components for Discrimination*) [158].

Um forma possível de extrair componentes PCD é utilizar um classificador MLP [4] de duas camadas, sujeito a um processo de treinamento incremental, onde neurônios são inseridos na camada intermediária ao longo do processo de treinamento. Os pesos que conectam estes neurônios aos nós de entrada da rede fornecerão, ao final do processo de extração, as componentes PCD.

Este processo de treinamento é iniciado considerando uma rede de N nós de entrada, 1 neurônio na camada intermediária e K neurônios de saída [3], onde todos neurônios possuem a tangente hiperbólica como função de ativação. Esta rede é treinada na forma usual, ou seja, visando a minimização do erro médio quadrático cometido em sua saída. Após a convergência, tem-se uma estimativa da primeira componente principal de discriminação, a qual, através de um processo de atuação conjunta com os pesos da camada de saída, é a direção de projeção que melhor promove a separação dos dados nas diferentes classes. Em seguida, mais um neurônio é inserido na camada intermediária, sendo os vetores de pesos associados às componentes já extraídas mantidos fixos por todo restante do processo de extração. Este processo é repetido por um total de p vezes, produzindo p neurônios na camada escondida da rede, onde p é o número de componentes PCD a serem extraídas. Trata-se portanto de um processo de treinamento construtivo e cooperativo, sendo esta metodologia também conhecida como PCD construtiva [2]. Com o objetivo de evitar mínimos locais [4] é adequado, no processo de extração de cada compo-

[3] É evidente que o valor de K é dependente do tipo de codificação adotada para as classes. Neste trabalho, considera-se uma codificação maximamente esparsa [4], ou seja, cada neurônio é responsável por uma classe, sendo treinado para apresentar o valor alvo +1,0 para eventos a ela pertencentes; e -1,0, em caso contrário.

nente, realizar várias tentativas, onde, para cada uma, uma nova inicialização dos pesos em treinamento é realizada. Uma vez concluídas todas tentativas, seleciona-se aquela que apresentar melhor eficiência de generalização, dando-se prosseguimento ao processo de extração.

Quando aplicada para fins de compactação, a análise PCD, por reter informação mais diretamente relacionada à classificação, em geral, produz sistemas de classificação baseados nas características extraídas que são mais eficientes que a análise PCA, ou seja, que obtêm uma eficiência de classificação similar ou superior, utilizando, em geral, menor número de componentes [5].

É interessante ainda observar que, dada a estrutura e a forma como a rede extratora PCD é treinada, as eficiências obtidas ao longo do processo de extração fornecem uma estimativa aproximada das eficiências de um classificador neural baseado em eventos compactados com base nas direções fornecidas por esta análise. Este resultado é similar ao obtido pela análise PCA com respeito aos autovalores, que estão diretamente relacionados ao erro quadrático médio cometido na compactação dos eventos, conforme a Equação B.13.

Bibliografia

[1] THEODORIDIS, S., KOUTROUMBAS, K., *Pattern Recognition*. 2 ed. Elsevier academic press, 2003.

[2] SOARES FILHO, W., *Classificação do Ruído Irradiado por Navios usando Redes Neurais*. Ph.D. dissertation, UFRJ, 2001.

[3] SOUZA FILHO, J., SEIXAS, J., "Classificador Neural Online para Sonar Passivo utilizando um Processador Digital de Sinais de Alto-Desempenho". In: *V CBRN - V Congresso Brasileiro de Redes Neurais*, 2001.

[4] HAYKIN, S., *Neural Networks: a Comprehensive Foundation*. 2 ed. Prentice-Hall, 1999.

[5] SOARES FILHO, W., SEIXAS, J. M., CALôBA, L. P., "Principal Component Analysis for Classifying Passive Sonar Signals". In: *IEEE International Symposium on Circuits and Systems*, pp. 1–4, 2001.

[6] TUFTS, D. W., IANNIELLO, J., LOURTIE, I., *et al.*, "The Past, Present and Future of Underwater Acoustic Signal Processing", *IEEE Signal Processing Magazine*, v. 15, n. 4, pp. 21–51, 1998.

[7] HALEY, T., "Applying Neural Networks to Automatic Active Sonar Classification". In: IEEE (ed.), *Proceedings of 10th International Conference on Pattern Recognition*, v. II, pp. 41–44, 1990.

[8] SOLINSKI, J., NASH, E. A., "Neural-Network Performance Assessment in Sonar Applications". In: *IEEE Conference on Neural Networks for Ocean Engineering*, pp. 1–12, 1991.

[9] HEMMINGER, T. L., PAO, Y.-H., "Detection and Classification of Underwater Acoustic Transients Using Neural Networks", *IEEE Transactions on Neural Networks*, v. 5, n. 5, pp. 712–718, 1994.

[10] KOBUS, D., RUSSOTI, J., SCHILICHTING, C., *et al.*, "Multimodal Detection and Recognition Performance of Sonar Operators", *Human Factors*, v. 28, n. 1, pp. 23–29, 1886.

[11] GHOSH, J., DEUSER, L. M., BECK, S. D., "A Neural Network Based Hybrid System for Detection, Characterization and Classification of Short-Duration Ocean Signals", *IEEE Journal of Oceanic Engineering*, v. 17, n. 4, pp. 351–363, 1992.

[12] PAPOULIS, A., *Probability, Random Variables and Stochastic Processes*. 4 ed. McGraw-Hill, 2002.

[13] LOURENS, J., "Passive Sonar Detection of Ships with Spectrograms". In: IEEE (ed.), *Proceedings on Communications and Signal Processing*, pp. 147–151, 1990.

[14] ROTH, M. W., "Survey of Neural Network Technology for Automatic Target Recognition", *IEEE Transactions on Neural Networks*, v. 1, n. 1, pp. 28–43, 1990.

[15] SOUZA FILHO, J., SEIXAS, J., "Implementação de Classificadores Neuronais na Tecnologia de Processadores Digitais de Sinais", *IV CBRN - IV Congresso Brasileiro de Redes Neurais*, pp. 390–395, 1999.

[16] HUANG, W., LIPPMANN, R., "Neural Network and Traditional Classifiers", *Neural Information Processing Systems*, pp. 387–396, 1987.

[17] GHOST, J., HWANG, K., "Mapping Neural Networks onto Mensage-passing Multicomputers", *Journal of parallel and distributed computing*, v. 6, pp. 291–330, 1989.

[18] GORMAN, R., "Neural Networks and the Classification of Complex Sonar Signals". In: *IEEE Conference on Neural Networks for Ocean Eigeneering*, pp. 283–290, 1991.

[19] GORMAN, R., SEJNOWSKI, T., "Learned Classification of Sonar Targets Using a Massively Parallel Network", *IEEE Transactions on Acoustics, Speach and Signal Processing*, v. 7, n. 7, pp. 1135–1140, 1988.

[20] CASSELMAN, F., FREEMAN, D., KERRIGAN, D., *et al.*, "A Neural Network-based Passive Sonar Detection and Classification Design with a Low False Alarm Rate". In: *Procedings of the International Conference on Neural Networks for Ocean Eigeneering*, pp. 49–55, 1991.

[21] WEBER, D., KRüGER, C., "Detection of Tonals in Lofargrams using Connectionist Methods". In: *Procedings of the International Conference on Neural Networks*, v. 31, pp. 1662–1666, 1993.

[22] WARD, M., STEVENSON, M., "Sonar Signal Detection and Classification using Artificial Neural Networks". In: *Conference on Eletrical and Computer Engineering*, v. 2, pp. 717–721, Canada, 2000.

[23] OPPENHEIM, A., *Discrete-Time Signal Processing*. 2 ed. Prentice Hall, 1999.

[24] VETTERLI, M., KOVACEVIC, J., *Wavelets and Subband Coding*. 1 ed. Prentice Hall, 1995.

[25] RUSSO, A. P., "Constrained Neural Networks for Recognition of Passive Sonar Signals Using Shape". In: *IEEE Conference on Neural Networks for Ocean Eigeneering*, pp. 69–76, 1991.

[26] AZIMI-SADJADI, M., YAO, D., HUANG, Q., *et al.*, "Underwater Target Classification Using Wavelet Packets and Neural Networks", *IEEE Transactions on Neural Networks*, v. 11, n. 3, pp. 784–794, 2000.

[27] VAN TREES, H. L., *Detection, Estimation and Modulation Theory*, part I and III. 1 ed. John Wiley Sons, 1968.

[28] RUMELHART, D., HINTON, G., WILLIAMS, R., *Learning Internal Representations by Error Propagation*, v. 1: foundations, parallel distributed processing: exploration in the microstructure of cognition ed., MIT Press, pp. 318–362, 1986.

[29] KHOTANZAD, A., LU, J., SRINATH, M., "Target Detection using a Neural Network based Passive Sonar System". In: *IEEE International Joint Conference on Neural Networks*, pp. 335–340, 1989.

[30] HOPFIELD, J., "Neural Network and Physical Systems with Emergent Collective Computational Abilities". In: *Procedings of the National Academy of Science*, v. 79, pp. 2554–2558, 1982.

[31] VAN-HOUTTE, P., DEEGAN, K., KHORASANI, K., "Passive Sonar Processing Using Neural Networks". In: *IEEE International Joint Conference on Neural Networks*, v. 2, pp. 1154–1159, 1991.

[32] WANG, Y., CRUZ, J., MULLIGAN, J., "Two Coding Strategies for Bidirecional Associative Memory", *IEEE Transactions on Neural Networks*, v. 1, n. 1, pp. 81–91, 1990.

[33] MOORE, P., ROITBLAT, H., PENNER, R., *et al.*, "Recognizing Successive Dolphin Echoes with an Integrator Gateway Technology", *Neural Networks*, v. 4, n. 6, pp. 701–709, 1991.

[34] FREEMAN, H., "Computer Processing of Line Drawing Images", *Computer Surveys*, v. 6, n. 1, 1974.

[35] "Defense Advanced Research Projects Agency", `http://www.darpa.mil`.

[36] LIPPMANN, R., "Pattern Classification using Neural Networks", *IEEE Communications Magazine*, v. 27, n. 11, pp. 47–50,59–64, 1989.

[37] OJA, E., "Principal Components, Minor Components and Linear Neural Networks", *Neural Networks*, v. 5, n. 6, pp. 927–935, 1992.

[38] SANGER, T., "Optimal Unsupervised Learning in a Single Linear Feedfoward Network", *Neural Networks*, v. 2, n. 6, pp. 459–473, 1989.

[39] KOHONEN, T., *Self-organization and Associative Memory*. 3 ed. Springer-Verlag, 1989.

[40] KOWALSKI, J., HARTMAN, E., KEELER, J., "Layered Neural Networks with Gaussian Hidden Units as Universal Approximators", *Neural Computation*, v. 2, pp. 210–215, 1990.

[41] SHIN, Y., GHOSH, J., "The PI-SIGMA Network: An Efficient Higher-order Network for Pattern Classification and Function Approximation". In: *Procedings of the International Joint Conference on Neural Networks*, v. 1, pp. 13–18, 1991.

[42] GROSSBERG, S., "Contour Enhancement, Short-term Memory and Constancies in Reverberating Networks", *Studies in Applied Mathematics*, v. 52, pp. 217–257, 1973.

[43] PAO, Y., *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley, 1989.

[44] DINIZ, P. S. R., DA SILVA, E. A. B., LIMA NETTO, S., *Digital Signal Processing: System Analysis and Design*. Cambridge University Press, 2002.

[45] MARPLE, S. L., *Digital Spectral Analysis with Applications*. Prentice-Hall, 1987.

[46] BARNSLEY, M., *Fractals Everywhere*. 3 ed. Academic Press, 1988.

[47] STEVENSON, M., *Survey of Dynamic Neural Networks Techniques with Applications to Temporal Processing Tasks*, Report, DREA, 1994.

[48] "Defense Research and Development website", `http://www.dreo.dnd.ca/`.

[49] DUDA, R. O., HART, P. E., STORK, D. G., *Pattern Classification*. John Willey & Sons, 2001.

[50] RABINER, L., A, B, *et al.*, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[51] WAN, C., GOH, J., CHEE, H., "Optimal Tonal Detectors Based on the Power Spectrum", *IEEE Jounal of Ocean Eigineering*, v. 25, n. 4, pp. 540–552, 2000.

[52] SHANMUGAN, K., BREIPOHL, A., *Random Signals - Detection, Estimation and Data Analysis*. John Wiley Sons, 1988.

[53] URICK, R., "Models for the Amplitude Fluctuations of Narrow-band Signals and Noise in the Sea", *Journal of the Acoustical Society of America*, v. 62, pp. 878–887, 1977.

[54] KRAMER, M., "Nonlinear Principal Component Analysis using Autoassociative Neural Networks", *AICHE Journal*, v. 37, pp. 365–375, 1991.

[55] SEIXAS, J. M., SOARES FILHO, W., CALôBA, L. P., "Compact On-line neural system for classifying passive sonar signals". In: *International Conference on Signal Processing, Applications & Technology*, pp. 1–4, 1999.

[56] CARPENTER, G., GROSSBERG, S., "ART2: Self-organization of Stable Category Recognition Codes for Analog Input Patterns", *Applied Optics*, v. 26, pp. 4919–4930, 1987.

[57] FERNANDES, H. L., *Classificação de Navios Baseada em Curvas Principais*. M.Sc. dissertation, UFRJ, 2005.

[58] HASTIE, T., STUEZLE, W., "Principal Curves", *Journal of the American Statistical Association*, v. 84, pp. 502–516, 1989.

[59] VERBEEK, J. J., VLASSIS, N., KRÖSE, B., "A k-segments Algorithm for Finding Principal Curves", *Pattern Recognition Letters*, v. 23, n. 8, pp. 1009–1017, 2002.

[60] NIELSEN, R. O., *Sonar Signal Processing*. Artech House, 1991.

[61] OJA, E., "A Simplified Neuron Model as a Principal Component Analyser", *Journal of Mathematical Biology*, v. 15, n. 3, pp. 267–273, 1982.

[62] KUNG, S., DIAMANTARAS, K., TAUR, J., "Adaptative Principal Component Extraction (APEX) and Applications", *Signal Processing*, v. 42, n. 5, pp. 1202–1217, 1994.

[63] COMON, P., GOLUB, G., "Tracking a Few Extreme Singular Values and Vectors in Signal Processing", *Proceedings of the IEEE*, v. 78, n. 8, pp. 1327–1343, 1990.

[64] DIAMANTARAS, K., KUNG, S., *Principal Component Neural Networks - Theory and Applications*. John Wiley and Sons, 1996.

[65] BALDI, P. F., HORNIK, K., "Learning in Linear Neural Networks: A Survey", *IEEE Transactions on Neural Networks*, v. 6, n. 4, pp. 837–858, 1995.

[66] HYVÄRINEN, J., KARHUNEN, J., OJA, E., *Independent Component Analysis*. John Willey & Sons, Inc., 2001.

[67] CICHOCKI, A., AMARI, S., *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley Sons, 2002.

[68] FIORI, S., PIAZZA, F., "A Comparison of Three PCA Neural Techniques". In: *Proc. of European Symposium on Artificial Neural Networks (ESANN)*, pp. 275–280, 1999.

[69] YANG, B., "Projection Approximation Subspace Tracking - PAST", *IEEE Transactions on Signal Processing*, v. 43, n. 1, pp. 95–107, 1995.

[70] STRANG, G., *Linear Algebra and Its Applications*. Saunders, 1988.

[71] FLETCHER, R., *Pratical Methods of Optimization - Volume 1 - Unconstrained Optimization*. John Wiley Sons, 1980.

[72] OJA, E., "Principal Components, Minor Components and Linear Neural Networks", *Neural Networks*, v. 5, n. 6, pp. 927–935, 1992.

[73] BROCKETT, R. W., "Least Squares Matching Problems", *Linear Algebra and Its Applications*, v. 122-124, pp. 761–777, 1989.

[74] BROCKETT, R., "Dynamical Systems that Sort Lists, Diagonalize Matrices and Solve Linear Programming Problems", *Linear algebra and its applications*, v. 146, pp. 79–91, 1991.

[75] XU, L., "Least Mean Square Error Reconstruction Principle for Self-Organizing Neural-Nets", *Neural Networks*, v. 6, pp. 627–648, 1993.

[76] DINIZ, P., *Adaptative Filtering - Algorithms and Pratical Implementation*. Kluwer Academic Publishers, 1997.

[77] FLETCHER, R., *Pratical Methods of Optimization - Volume 2 - Constrained Optimization*. John Wiley Sons, 1980.

[78] CHATTERJEE, C., KANG, Z., ROYCHOWDHURY, V., "Algorithms for Accelerated Convergence of Adaptive PCA", *IEEE Transactions on Neural Networks*, v. 11, n. 2, pp. 338–355, 2000.

[79] KARHUNEN, J., JOUTSENSALO, J., "Generalizations of Principal Component Analysis, Optimizations Problems, and Neural Networks", *Neural Networks*, v. 8, n. 4, pp. 549–562, 1995.

[80] MAO, Y., HUA, Y., "Fast Subspace Tracking and Neural Network Learning by a Novel Information Criterion", *IEEE Transactions on Signal Processing*, v. 46, n. 7, pp. 1967–1979, 1998.

[81] OUYANG, S., BAO, Z., "Fast Principal Component Extraction by a Weighted Information Criterion", *IEEE Transactions on Signal Processing*, v. 50, n. 8, pp. 1994–2002, 2002.

[82] BANNOUR, S., AZIMI-SADJADI, M. R., "Principal Component Extraction using Recursive Least Squares Learning", *IEEE Transactions on Neural Networks*, v. 6, n. 2, pp. 457–469, 1995.

[83] ABED-MERAIM, K., CHKEIF, A., HUA, Y., "Fast Orthonormal PAST Algorithm", *IEEE Signal Processing Letters*, v. 7, n. 3, pp. 60–62, 2000.

[84] BOURLAND, H., KAMP, Y., "Auto-association by Multilayer Perceptrons and Singular Value Decomposition", *Biological Cybernetics*, v. 59, pp. 291–294, 1988.

[85] BALDI, P., HORNIK, K., "Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima", *Neural Networks*, v. 2, n. 1, pp. 53–58, 1988.

[86] SOUZA FILHO, J. B. O., SOUZA, M., CALôBA, L., *et al.*, "Extração de Componentes Principais Lineares em Aplicações de Alta-dimensionalidade", *V SBAI - Simpósio Brasileiro de Automação Inteligente*, , 2001.

[87] DIAMANTARAS, K., KUNG, S., "Multilayer Neural Networks for Reduced-rank Aproximation", *IEEE Transactions on neural networks*, v. 5, n. 5, pp. 684–697, 1994.

[88] GOLUB, G., LOAN, C., *Matrix Computations*. The Johns Hopkins University Press, 1989.

[89] AMARI, S., "Neural Theory of Association and Concept Formation", *Biological Cybernetics*, v. 79, n. 8, pp. 175–185, 1977.

[90] OJA, E., KARHUNEN, J., "On Stochastic Approximation of the Eigenvectors and Eigenvalues of the Expectation of a Random Matrix", *Journal of Mathematical Analysis and Applications*, v. 106, pp. 69–54, 1985.

[91] OJA, E., OGAWA, H., WANGVIWATTANA, J., "Principal Component Analysis by Homogeneous Neural Networks. Part I: The Weighted Subspace Criterion", *IEICE Transactions on Information and Systems*, v. 3, pp. 366–375, 1992.

[92] OJA, E., OGAWA, H., WANGVIWATTANA, J., "Principal Component Analysis by Homogeneous Neural Networks. Part II: Analysis and Extensions of the Learning Algorithms", *IEICE Transactions on Information and Systems*, v. 3, pp. 376–382, 1992.

[93] RUBNER, J., TAVAN, P., "A Self-organizing Network for Principal Component Analysis", *Europhysics Letters*, v. 10, pp. 693–698, 1989.

[94] FIORI, S., PIAZZA, F., "A General Class of ψ-APEX Neural Algorithms", *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, v. 47, n. 9, pp. 1394–1397, 2000.

[95] CICHOCKI, A., KASPRZAK, W., SKARBEK, W., "Adaptive Learning Algorithm for Principal Component Analysis with Partial Data", *Austrian Society for Cybernetic Studies*, v. 2, pp. 1014–1019, 1996.

[96] SOUZA FILHO, J. B. O., CALÔBA, L. P., SEIXAS, J. M., "An Accurate and Fast Neural Method for PCA Extraction". In: *International Joint Conference on Neural Networks*, pp. 586–591, 2004.

[97] SOUZA FILHO, J. B. O., *Análise de Componentes Principais em Sistemas de Sonar*. M.Sc. dissertation, Universidade Federal do Rio de Janeiro - UFRJ, 2002.

[98] BOULARD, H., KAMP, Y., "Auto-association by Multilayer Perceptrons and Singular Value Decomposition", *Biological Cybernetics*, v. 59, n. 4-5, pp. 291–294, 1988.

[99] HAYKIN, S., *Adaptive Filter Theory*. Prentice-Hall International Editions, 1991.

[100] MAO, J., JAIN, A., "Artificial Neural Networks for Feature Extraction and Multivariate Data Projection", *IEEE Transactions on Neural Networks*, v. 6, n. 2, pp. 296–317, 1995.

[101] BORAY, G. K., SRINATH, M. D., "Conjugate Gradient Techniques for Adaptive Filtering", *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, v. 39, n. 1, pp. 1–10, 1992.

[102] MOLLER, M., "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning", *Neural Networks*, v. 6, n. 4, pp. 525–533, 1993.

[103] HESTENES, M. R., STIEFEL, S., "Methods of Conjugate Gradient for Solving Linear Systems", *Journal of Research of the National Bureau of Standards*, v. 49, n. 6, pp. 409–436, 1952.

[104] SHEWCHUK, J. R., "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain", `http://www.cs.cmu.edu/~jrs/jrspapers.html`, 1994.

[105] YANG, X., SARKAR, T. K., ARVAS, E., "A Survey of Conjugate Gradient Algorithms for Solution of Extreme Eigen-problems of a Symmetric Matrix", *IEEE Transactions on Acoustics, Speach and Signal Processing*, v. 37, n. 10, pp. 1550–1556, 1989.

[106] FU, Z., DOWNLING, E. M., "Conjugate Gradient Eigenstructure Tracking for Adaptive Spectral Estimation", *IEEE Transactions on Signal Processing*, v. 43, n. 5, pp. 1151, 1995.

[107] FLETCHER, R., REEVES, C. M., "Function Minimization by Conjugate Gradients", *The Computer Journal*, v. 7, pp. 149–154, 1964.

[108] POLAK, E., *Computational Methods in Optimization*. New York: Academic Press, 1971.

[109] POWELL, M. J. D., "Restart Procedures for the Conjugate Gradient Method", *Mathematical Programming*, v. 12, pp. 241–254, 1977.

[110] RIEDMILLER, M., BRAUN, H., "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm". In: *IEEE Conference on Neural Networks*, pp. 586–591, 1993.

[111] RIEDMILLER, M., "Advanced Supervised Learning in Multilayer Perceptrons - from Backpropagation to Adaptive Learning Algorithms", *Computer Standards and Interfaces*, v. 16, pp. 265–278, 1994.

[112] IGEL, C., HUSKEN, M., "Empirical Evaluation of the Improved RPROP Learning Algorithms", *Neurocomputing*, v. 50, pp. 105–123, 2003.

[113] IGEL, C., HUSKEN, M., "Improving the RPROP Learning Algorithm". In: *Second International Symposium on Neural Computation*, pp. 115–121, 2000.

[114] SOUZA FILHO, J. B. O., SEIXAS, J. M., "IR-PCA: Um Algoritmo Neural Acurado e Rápido para a Extração de Componentes Principais". In: *XVI CBA - Congresso Brasileiro de Automática*, 2006.

[115] PRESS, W., VATTERLING, W., TEUKOLSY, S., *et al.*, *Numerical Recipies in C - The Art of Scientific Computing*. Cambrige University Press, 1995.

[116] AKIPPI, C., BRAIONE, P., "Classification Methods, Reduced Datasets and Quality Analysis Applications". In: *IEEE International Conference on Computational Intelligence for Mesurement Systems and Applications*, pp. 121–126, 2004.

[117] MEDEIROS, M. C., TERÄSVIRTA, T., RECH, G., "Building Neural Network Models for Time Series: A Statistical Approach", *Journal of Forecasting*, v. 25, pp. 49–75, 2006.

[118] GERMAN, S., BIENENSTOCK, E., DOURSAT, R., "Neural Networks and Bias/Variance Dilemma", *Neural Computation*, v. 4, n. 1, pp. 1–58, 1992.

[119] LEVIN, A. U., LEEN, T. K., MOODY, J. E., "Fast Pruning Using Principal Components". In: *6th NIPS - Neural Information Processing Systems Conference*, v. 6, pp. 35–42, 1994.

[120] CUN, Y. L., DENKER, J. S., SOLLA, S. A., "Optimal Brain Damage". In: *2th NIPS - Neural Information Processing Systems Conference*, v. 2, pp. 35–42, 1990.

[121] GUYON, I., VAPNIK, V., BOSER, B., *et al.*, "Structural Risk Minimization for Character Recognition". In: *4th NIPS - Neural Information Processing Systems Conference*, v. 2, pp. 471–479, 1992.

[122] GIROSI, F., JONES, M., POGGIO, T., "Regularization Theory and Neural Networks Architectures", *Neural Computation*, v. 7, n. 2, pp. 219–269, 1995.

[123] PAREKH, R., YANG, J., HONAVAR, V., "Constructive Neural-Network Learning Algorithms for Pattern Recognition", *IEEE Transactions on Neural Networks*, v. 11, n. 2, pp. 436–451, 2000.

[124] ANDERS, U., KORN, O., "Model Selection in Neural Networks", *Neural Networks*, v. 12, n. 2, pp. 309–323, 1999.

[125] MEDEIROS, M., TERäSVIRTA, T., RECH, G., "Building Neural Networks Models for Time Series: A Statistical Approach", *Journal of Forecasting*, v. 25, n. 1, pp. 49–75, 2006.

[126] MURATA, N., YOSHIZAWA, S., AMARI, S., "A Criterion for Determining the Number of Parameters in an Artificial Neural Network Model", *IEEE Transactions on Neural Networks*, v. 5, n. 6, pp. 865–872, 1994.

[127] ALIPPI, C., "Selecting Accurate, Robust and Minimal Feedfoward Neural Networks", *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, v. 49, n. 12, pp. 1799–1810, 2002.

[128] VÁZQUEZ, G., P, G. R. L., JOAQUíN, P. J., *et al.*, "Model Selection Methods in Multilayer Perceptrons". In: *IEEE International Joint Conference on Neural Networks*, v. 2, pp. 25–29, 2004.

[129] STONE, M., "Cross-validatory Choice and Assessement of Statistical Predictions", *Jounal of the Royal Statistical Society*, v. 36, n. 2, pp. 111–147, 1974.

[130] PRECHELT, L., "Automatic Early Stopping using Cross-validation: Quantifying the Criteria", *Neural Networks*, v. 11, n. 4, pp. 761–766, 1998.

[131] FINNOFF, W., HERGERT, F., ZIMMERMANN, H. G., "Improving Model Selection by Nonconvergent Methods", *Neural Networks*, v. 6, n. 6, pp. 771–783, 1993.

[132] KAY, S. M., *Fundamentals of Statistical Signal Processing: Estimation Theory*, v. I: Estimation Theory. Prentice Hall, 1993.

[133] SEIXAS, J. M., DAMAZIO, D. O., "A Neural Particle Discriminator For Calorimetry in High Energy Physics". In: *Terceiro Congresso Brasileiro de Redes Neuronais*, Florianópolis, 1997.

[134] ABRAMOWITZ, M., STEGUN, I. A., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S Department of Commerce, 2002.

[135] HANLEY, J. A., MCNEIL, B. J., "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve", *Diagnostic Radiology*, v. 143, n. 1, pp. 29–36, 1982.

[136] SRINIVASAN, A., *Note on the Location of Optimal Classifiers in N-Dimensional ROC Space*, Report, Oxford University, 1999.

[137] EDWARDS, D. C., METZ, C. E., NISHIKAWA, R. M., "The Hypervolume under the ROC Hypersurface of "Near-Guessing" and "Near-Perfect" observers

in N-Class classification tasks", *IEEE Transactions on Medical Imaging*, v. 24, n. 3, pp. 293–299, 2005.

[138] LACHICHE, N., FLACH, P. A., "Improving Accuracy and Cost of Two-class and Multi-class Probabilistic Classifiers using ROC Curves". In: *Proc. 20th International Conference on Machine Learning (ICML'03)*, pp. 416–423, 2003.

[139] FAWCETT, T., "Using Rule Sets to Maximize ROC Performance". In: *IEEE International Conference on Data Mining (ICDM-01)*, pp. 131–138, 2001.

[140] HAND, D. J., TILL, R. J., "A Simple Generalization of the Area Under the ROC Curve for Multiple Class Classification Problems", *Machine Learning*, v. 45, n. 2, pp. 171–186, 2001.

[141] DYBOWSKI, R., GANT, V. (eds.), *Clinical Applications of Artificial Neural Networks*. Cambridge University Press, 2001.

[142] PENNY, W. D., ROBERTS, S. J., *Neural Network Predictions with Error Bars*, Report, Department of Eletrical and Eletronic Engineering, Imperial College, London, 1997.

[143] TIBSHIRANI, R., "A Comparison of Some Error Estimates for Neural Networks Models", *Neural Computation*, v. 8, n. 1, pp. 152–163, 1996.

[144] KOHAVI, R., "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection". In: *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1145, 1995.

[145] ZANG, P., "Model Selection via Multifold Cross-validation", *Annals of Statistics*, v. 21, n. 2, pp. 299–313, 1993.

[146] MONSTELLER, F., TUKEY, J. W., *Handbook of Social Psychology - Data analysis, including Statistics*, v. 2. Addison-Wesley, 1968.

[147] EFRON, B., "Another Look at the Jackknife", *Annals of Statistics*, v. 7, pp. 1–26, 1979.

[148] EFRON, B., TIBSHIRANI, R., *An Introduction to the Boostrap*. Chapman and Hall, 1993.

[149] EFRON, B., "Estimating the Error Rate of a Prediction Rule: Some Improvements on Cross-validation", *Journal of the American Statistical Association*, v. 78, pp. 316–331, 1983.

[150] EFRON, B., TIBSHIRANI, R., "Improvements on Cross-validation: the 0.632+ Bootstrap Method", *Journal of the American Statistical Association*, v. 92, pp. 540–560, 1997.

[151] CHERNICK, M. R., *Bootstrap Methods - A Practioner´s Guide*. Wiley Series in Probability and Statistics, 1999.

[152] ZOUBIR, A., BOASHASH, B., "The Bootstrap and Its Application in Signal Processing", *IEEE Signal Processing Magazine*, v. 15, n. 1, pp. 56–76, 1998.

[153] NADEAU, C., BENGIO, Y., "Inference for the Generalization Error", *Machine Learning*, v. 52, n. 3, pp. 239–281, 2003.

[154] BENGIO, Y., GRANDVALET, Y., "No Unbiased Estimator of the Variance of k-Fold Cross-validation", *Journal of Machine Learning Research*, v. 5, pp. 1089–1105, 2004.

[155] DIETTERICH, T. G., "Approximative Statistical Tests for Comparing Supervised Classification Learning Algorithms", *Neural Computation*, v. 10, n. 7, pp. 1895–1923, 1998.

[156] JAIN, A. K., MURTY, M. N., FLYNN, P. J., "Data Clustering: a Review", *ACM Computing Surveys*, v. 31, n. 3, pp. 264–323, 1999.

[157] JOLLIFFE, I. T., *Principal Component Analysis*. 2 ed. Springer-Verlag, 2002.

[158] CALôBA, L., PEREIRA, F. S., SEIXAS, J., "Neural Discriminating Analysis for a Second-Level Trigger System". In: *International Conference on Computing in High Energy Physics*, Rio de Janeiro, RJ, Brasil, 1995.

[159] DUNN, J. C., "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-separated Clusters", *Journal of Cybernetics*, v. 3, pp. 32–57, 1973.

[160] DAVIES, D. L., BOULDIN, D. W., "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis on Machine Intelligence*, v. 1, n. 2, pp. 224–227, 1979.

[161] HUBERT, L., ARABIE, P., "Comparing Partitions", *Journal of Classification*, v. 2, n. 1, pp. 193–218, 1985.

[162] BEZDEK, J. C., PAL, N. R., "Some New Indexes of Cluster Validity", *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, v. 28, n. 3, pp. 301–314, 1998.

[163] KAUFMAN, L., ROUSSEEUW, P. J. (eds.), *Finding Groups in Data - An Introduction to Cluster Analysis*. 2 ed. John Wiley & Sons, 2005.

[164] VASSALI, M. R., SEIXAS, J. M., CALôBA, L. P., "A Neural Particle Discriminator Based on a Modified ART Architecture". In: *IEEE International Symposium on Circuits and Systems*, v. II, pp. 121–124, 2002.

[165] MILLIGAN, G. W., COPPER, M. C., "An Examination Procedures for Determining the Number of Clusters in a Dataset", *Psychometrika*, v. 50, n. 2, pp. 159–179, 1985.

[166] TIBSHIRANI, R., WALTHER, G., HASTIE, T., *Estimating the Number of Clusters in a Dataset via the Gap Statistic.*, Report, Departament of Statistics, Stanford University, 2000.

[167] MONTI, S., TAMAYO, P., MESIROV, J., *et al.*, "Consensus Clustering: a Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data", *Machine learning*, v. 52, n. 1-2, pp. 91–118, 2003.

[168] TIBSHIRANI, R., WALTHER, G., *Cluster Validation by Prediction Strengh*, Report, Departament of Biostatistics, Stanford University, 2001.

[169] ZHANG, J., MODESTINO, W., "A Model-fitting Approach to Cluster Validation with Application to Stochastic Model-based Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 12, n. 10, pp. 1009–1017, 1990.

[170] GOWER, J., "A Comparison of Some Methods of Cluster Analysis", *Biometrics*, v. 23, n. 4, pp. 623–628, 1967.

[171] FLOREK, K., PERKAL, J., STEINHAUS, H., *et al.*, "Sur La Liaison et La Division des Points d´un Ensemble Fin", *Colloquium Mathematicum*, v. 2, pp. 282–285, 1951.

[172] SNEATH, P. H. A., "Some Thoughts on Bacterial Classification", *Journal of General Microbiology*, v. 17, n. 1, pp. 184–200, 1957.

[173] MCQUITTY, L. L., "Hierarquical Linkage Analysis for the Isolation of Types", *Educational and Psychological Measurement*, v. 20, pp. 55–67, 1960.

[174] SOKAL, R. R., SNEATH, P. H. A., *Principles of Numerical Taxonomy*. San Francisco, Freeman, 1963.

[175] SOKAL, R. R., MICHENER, C. D., "A Statistical Method for Evaluating Systematic Relationships", *University of Kansas Science Bulletin*, v. 38, pp. 1409–1438, 1958.

[176] WARD, J. H. J., "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, v. 58, n. 301, pp. 236–244, 1963.

[177] BOBERG, J., SALAKOSKI, T., "General Formulation and Evaluation of Agglomerative Clustering Methods with Metric and Non-metric Distances", *Pattern Recognition*, v. 26, n. 9, pp. 1395–1406, 1993.

[178] HALKIDI, M., BATISTAKIS, Y., VAZIRGIANNIS, M., "On Clustering Validation Techniques", *Journal of Intelligent Information Systems*, v. 17, n. 2-3, pp. 107–145, 2001.

[179] ROUSSEEUW, P. J., "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Journal of Computational and Applied Mathematics*, v. 20, n. 1, pp. 53–65, 1987.

[180] PREPARATA, F., SHARMOS, M., *Computational Geometry: An Introduction*. Springer-Verlag, 1987.

[181] BOLSHAKOVA, N., AZUAJE, F., "Cluster Validation Techniques for Genome Expression Data", *Signal Processing*, v. 83, n. 4, pp. 825–833, 2003.

[182] "Physics Analysis Workstation", `http://paw.web.cern.ch/paw/`.

[183] RAO, C. R., "The Utilization of Multiple Measurements in Problems of Biological Classification", *Journal of the Royal Statistical Society*, v. 10, n. 2, pp. 159–193, 1948.

[184] DICE, L. R., "Measures of the Amount of Ecologic Association Between Species", *Ecology*, v. 26, n. 3, pp. 297–302, 1945.

[185] HO, T. K., HULL, J. J., SRIHARI, S. N., "Decision Combination in Multiple Classifier Systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 16, n. 1, pp. 90–94, 1994.

[186] ERIC, R., GAWTHROP, P., *Modular Neural Networks: a State of the Art*, Report, Centre for System and Control. Faculty of mechanical Engineering, University of Glasgow, United Kingdom, 1995.

[187] HASHEM, S., "Optimal Linear Combination of Neural Networks", *Neural Networks*, v. 10, n. 4, pp. 599–614, 1997.

[188] HUANG, Y. S., SUEN, C. Y., "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 17, n. 1, pp. 90–94, 1995.

[189] AZAM, F., *Biologically Inspired Modular Neural Networks*. Ph.D. dissertation, Viginia Polytechnic Institute and State University, 2000.

[190] HANSEN, L. K., SALAMON, P., "Neural Network Ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 12, n. 10, pp. 993–1001, 1990.

[191] BENEDIKTSSON, J. A., SVEINSSON, J. R., ERSOY, O. K., *et al.*, "Parallel Consensual Neural Networks", *IEEE Transactions on Neural Networks*, v. 8, n. 1, pp. 54–64, 1997.

[192] SCHMIDT, A., *A Modular Neural Network Architecture with Additional Generalization Abilities for High Dimensional Input Vectors*. M.Sc. dissertation, Metropolitan University, 1996.

[193] BREIMAN, L., "Bagging Predictors", *Machine Learning*, v. 24, n. 2, pp. 123–140, 1996.

[194] SCHAPIRE, R. E., "The Strength of Weak Learnability", *Machine Learning*, v. 5, n. 2, pp. 197–227, 1990.

[195] FREND, Y., SCHAPIRE, R. E., "Experiments with a New Boosting Algorithm". In: *Proceedings of the Thirteenth International Conference in Machine Learning*, pp. 197–227, 1996.

[196] SHIMSHONI, Y., INTRATOR, N., "Classification of Seismic Signals by Integrating Ensembles of Neural Networks", *IEEE Transactions on Signal Processing*, v. 46, n. 5, pp. 1194–1201, 1998.

[197] ISLAN, M. M., YAO, X., MURASE, K., "A Constructive Algorithm for Training Cooperative Neural Networks Ensembles", *IEEE Transactions on Neural Networks*, v. 14, n. 4, pp. 820–834, 2003.

[198] KITTLER, J., HATEF, M., DUIN, R. P. W., *et al.*, "On Combining Classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, n. 3, pp. 226–239, 1998.

[199] XU, L., SUEN, C. Y., "Methods of Combining Multiple Classifiers and their Applications to Handwriting Recognition", *IEEE Transactions on Systems, Man and Cybernetics*, v. 22, n. 3, pp. 418–435, 1992.

[200] WOODS, K., KEGELMEYER, W. P., BOWYER, K., "Combination of Multiple Classifiers using Local Accuracy Estimates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 4, pp. 405–410, 1997.

[201] PERRONE, M. P., COOPER, L. N., "When Networks Disagree: Ensemble Methods for Hybrid Neural Networks". In: Mammone, R. J. (ed.), *Neural Networks for Speech and Image Processing*, Chapman-Hall, pp. 126–142, 1993.

[202] BREIMAN, L., "Stacked Regressions", *Machine Learning*, v. 24, n. 1, pp. 49–64, 1996.

[203] UEDA, N., "Optimal Linear Combination of Neural Networks for Improving Classification Performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 2, pp. 207–215, 2000.

[204] VAN ESSEN, D. C., ANDERSON, C. H., FELLMAN, D. J., "Information Processing in the Primate Visual System: an integrated systems perspective", *Science*, v. 255, n. 5043, pp. 419–423, 1992.

[205] GROSSBERG, S., "Adaptive Pattern Classification and Universal Recoding: I. Parallel Development and Coding of Neural Feature Detectors", *Biological Cybernetics*, v. 23, n. 3, pp. 121–134, 1976.

[206] CASTILLO, F., *Incremental Neural Networks*, Report, INPG, Grenoble, France, 1991.

[207] FRENCH, R. M., "Catastrophic Forgetting in Connectionist Networks", *Trends in Cognitive Sciences*, v. 3, n. 4, pp. 128–135, 1999.

[208] BARTFEI, G., "Hierarchical Clustering with Art Neural Networks". In: *World Congress on Computational Intelligence*, v. 2, pp. 940–944, Florida, USA, 1994.

[209] TSAI, H., TAI, H., REYNOLDS, A., "An ART2-BP Supervised Neural Net". In: *World Congress on Neural Networks*, v. 2, pp. 619–624, San Diego, USA, 1994.

[210] DE BOLLIVIER, M., GALLINARI, P., THIRIA, S., "Cooperation of Neural Nets and Task Decomposition". In: *International Joint Conference on Neural Networks*, v. 2, pp. 573–576, 1991.

[211] AUDA, G., KAMEL, M., RAAFAT, H., "Modular Neural Network Architecture for Classification". In: *IEEE International Joint Conference on Neural Networks*, v. 3, pp. 1240–1243, Perth, Australia, 1995.

[212] AUDA, G., KAMEL, M., RAAFAT, H., "Modular Neural Network Architecture for Classification". In: *IEEE International Joint Conference on Neural Networks*, v. 2, pp. 1279–1284, 1996.

[213] JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J., *et al.*, "Adaptive Mixtures of Local Experts", *Neural computation*, v. 3, n. 1, pp. 79–87, 1991.

[214] JORDAN, M. I., JACOBS, R. A., "Hierarchical Mixtures of Experts and EM Algorithm", *Neural computation*, v. 6, pp. 181–214, 1994.

[215] LIU, Y., YAO, X., "Evolving Modular Neural Networks Which Generalize Well". In: *Proceedings of the IEEE Conference on Evolutionary Computation*, pp. 605–610, 1997.

[216] CHO, S.-B., SHIMOHARA, K., "Evolutionary Learning of Modular Neural Networks with Genetic Programming", *Applied Intelligence*, v. 9, pp. 191–200, 1998.

[217] MERLIN, P., GONZALEZ, C., GONZALEZ, F., *et al.*, "Face Recognition Using Modular Neural Networks and Fuzzy Sugeno Integral for Response Integration". In: *Proceedings of the IEEE Joint Conference on Neural Networks*, pp. 349–354, Montreal, Canada, 2005.

[218] SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificação de Sinais de Sonar Passivo baseada em Filtragem Casada". In: *XV CBA - Congresso Brasileiro de Automática*, 2004.

[219] KÉGL, B., KRZYZAK, A., LINDER, T., *et al.*, "Learning and Design of Principal Curves", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 3, pp. 281–297, 2000.

[220] SANDILYA, S., KULKARNI, S. R., "Principal Curves with Bounded Turn", *IEEE Transactions on Information Theory*, v. 48, n. 10, pp. 2789–2793, 2002.

[221] TIBSHIRANI, R., "Principal Curves Revisited", *Statistics and Computation*, v. 2, pp. 183–190, 1992.

[222] SOUZA FILHO, J. B. O., SEIXAS, J. M., "Curvas Principais na Classificação de Sinais de Sonar Passivo". In: *II Workshop em Acústica Submarina*, 2002.

[223] FERNANDES, H. L., SOUZA FILHO, J. B. O., SEIXAS, J. M., "Classificação de Sinais Acústicos Submarinos Utilizando Curvas Principais". In: *XV CBA - Congresso Brasileiro de Automática*, 2004.

[224] ANAND, R., MEHROTRA, K., MOHAN, C. K., *et al.*, "Efficient Classification for Multi-class Problems Using Modular Neural Networks", *IEEE Transactions on Neural Networks*, v. 6, n. 1, pp. 117–124, 1995.

[225] HONG, T., FANG, T. C., HILDER, D., "PD Classification by a Modular Neural Network based on Task Decomposition", *IEEE Transactions on Dieletrics and Eletrical Insulation*, v. 3, n. 2, pp. 207–212, 1996.

[226] SINGH, S., MARKOU, M., "An Approach to Novelty Detection Applied to the Classification of Image Regions", *IEEE Transactions on Knowledge and Data Engineering*, v. 16, n. 4, pp. 396–407, 2004.

[227] SINGH, S., MARKOU, M., "A Neural Network-based Novelty Dectector for Image Sequence Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 28, n. 10, pp. 1664–1677, 2006.

[228] LU, B.-L., ITO, M., "Task Decomposition and Module Combination Based on Class Relations: A Modular Neural Network for Pattern Recognition", *IEEE Transactions on Neural Networks*, v. 10, n. 5, pp. 117–124, 1999.

[229] LU, B.-L., SHIN, J., ICHIKAWA, M., "Massively Parallel Classification of Single-Trial EEG Signals Using a Min-Max Modular Neural Network", *IEEE Transactions on Biomedical Engineering*, v. 51, n. 1, pp. 551–558, 2004.

[230] MA, Q., LU, B.-L., ISAHARA, H., "Part of Speach with Min-Max Modular Neural Networks". In: *IEEE International Conference on System, Man and Cybernetics*, v. 5, pp. 356–360, 1999.

[231] ANAND, R., MEHROTRA, K. G., MOHAN, C. K., *et al.*, "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets", *IEEE Transactions on Neural Networks*, v. 4, n. 6, pp. 962–968, 1993.

[232] ASH, T., "Dynamic Node Creation in Backpropagation Networks", *Connection Science*, v. 1, n. 4, pp. 365–375, 1989.

[233] MOODY, J., "Prediction Risk and Architecture Selection for Neural Networks". In: Cherkassky, V., Friedman, J. H., Wechsler, H. (eds.), *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, Springer, NATO ASI Series F, 1994.

[234] MARKOU, M., SINGH, S., "Novelty Detection: a Review - Part 1: Statistical Approaches", *Signal Processing*, v. 83, n. 12, pp. 2481–2497, 2003.

[235] TAX, D. M. J., DUIN, R. P. W., "Outlier Detection Using Classifier Instability". In: *Joint IAPR International Workshops in Advances in Pattern Recognition*, pp. 593–601, 1998.

[236] SAUNDERS, R., GERO, J., "The Importance of Being Emergent". In: *Proceedings of Artificial Intelligence in Design*, 2000.

[237] MARKOU, M., SINGH, S., "Novelty Detection: a Review - Part 2: Neural Network Based Approaches", *Signal Processing*, v. 83, n. 12, pp. 2499–2521, 2003.

[238] FUMERA, G., ROLI, F., GIACINTO, G., "Reject Option with Multiple Thresholds", *Pattern Recognition*, v. 33, n. 12, pp. 2099–2101, 2000.

[239] DESFORGES, M. J., JACOB, P. J., COOPER, J. E., "Applications of Probability Density Estimation to the Detection of Abnormal Conditions in Engineering". In: *Proceedings of Institute of Mechanical Engineers*, v. 212, 1998.

[240] ROBERTS, S. J., TARASSENKO, L., "A Probabilistic Resource Allocating Network for Novelty Detection", *Neural Computation*, v. 6, n. 2, pp. 270–284, 1994.

[241] TARASSENKO, L., "Novelty Detection for the Identification of Masses in Mammograms". In: *IEEE International Conference on Artificial Neural Networks*, v. 4, pp. 442–447, 1995.

[242] TARASSENKO, L., NAIRAC, A., TOWNSEND, N., *et al.*, "Novelty Detection in Jet Engines". In: *IEE Colloquium on Condition Monitoring, Imagery, External Structures and Health*, pp. 41–45, 1999.

[243] GUTTORMSSON, S. E., II, R. J. M., EL-SHARKAWI, M. A., "Elliptical Novelty Grouping for On-line Short-turn Detection of Excited Running Rotors", *IEEE Transactions on Energy Conversion*, v. 14, n. 1, pp. 16–22, 1999.

[244] ZHANG, B. T., VEENKER, G., "Neural Networks that Teach Themselves Through Genetic Discovery of Novel Examples". In: *Proceedings IEEE International Joint Conference on Neural Networks*, v. 1, pp. 690–695, 1991.

[245] VASCONCELOS, G. C., FAIRHURST, M. C., BISSET, D. L., "Investigating Feedforward Neural Networks with Respect to the Rejection of Spurious Patterns", *Pattern Recognition Letters*, v. 16, n. 2, pp. 207–212, 1995.

[246] WEI, F., MILLER, M., STOLFO, S. J., *et al.*, "Using Artificial Anomalies to Detect Unknown and Known Network Intrusions". In: *Proceedings of IEEE International Conference on Data Mining*, pp. 123–130, 2001.

[247] BOSER, Y. L. B., DENKER, J. S., HENDERSON, D., *et al.*, *Advances in Neural Information Processing Systems*, v. 2, chapter Handwritten Digit Recognition with a Back-propagation Network, Morgan Kaufman, pp. 396–404, 1990.

[248] CORDELLA, L. P., De Stefano, C., TORTORELLA, F., *et al.*, "A Method for Improving Classification Reliability of Multilayer Perceptrons", *IEEE Transactions on Neural Networks*, v. 6, n. 5, pp. 1140–1147, 1995.

[249] STEFANO, C. D., SANSONE, C., VENTO, M., "To Reject or Not to Reject: That is the Question - An Answer in Case of Neural Classifiers", *IEEE Transactions on Systems, Man and Cybernetics*, v. 30, n. 1, pp. 84–94, 2000.

[250] MOYA, M. R., KOCH, M. W., HOSTETLER, L. D., "One-class Classifier Networks for Target Recognition Applications". In: *Proceedings of World Congress on Neural Networks*, pp. 797–801, 1993.

[251] FREDRICKSON, S., ROBERTS, S., TOWNSEND, N., *et al.*, "Speaker Identification Using Networks of Radial Basis Functions". In: *Proceedings of the VII European Signal Processing Conference*, pp. 812–815, 1994.

[252] STREIFEL, R. J., MAKS, R. J., EL-SHARKAWI, M. A., "Detection of Shorted-turns in the Field of Turbine-generator Rotors Usign Novelty Dectectors - Development and Field Tests", *IEEE Transactions on Energy Conversarion*, v. 11, n. 2, pp. 312–317, 1996.

[253] WORDEN, K., "Structural Fault Detection Using a Novelty Measure", *Journal of Sound and Vibration*, v. 201, n. 1, pp. 85–101, 1997.

[254] SURACE, C., WORDEN, K., TOMLINSON, G., "A Novelty Detection Approach to Diagnose Damage in a Cracked Beam". In: *Proceedings of International Society of Optical Engineering*, v. 3089, pp. 947–953, 1997.

[255] HARRIS, T., "A Kohonen SOM Based, Machine Health Monitoring System which Enables Diagnosis of Faults Not Seen in the Training Set". In: *Proceedings of International Joint Conference on Neural Networks*, v. 1, pp. 25–29, 1993.

[256] LABIB, K., VEMURI, V. R., "NSOM: A Tool To Detect Denial of Service Attacks Using Self-Organizing Maps", http://citeseer.ist.psu.edu/labib03nsom.html.

[257] KWOK, T.-Y., YEUNG, D.-Y., "Objective Functions for Training New Hidden Units in Constructive Neural Networks", *IEEE Transactions on Neural Networks*, v. 8, n. 5, pp. 1131–1148, 1997.