Modelo semi-supervisionado aplicado à previsão da eficiência da Quimioterapia Neoadjuvante no tratamento de Câncer de Mama

Frederico Gualberto Ferreira Coelho

Modelo semi-supervisionado aplicado à previsão da eficiência da Quimioterapia Neoadjuvante no tratamento de Câncer de Mama

Frederico Gualberto Ferreira Coelho

Orientador: *Prof Dr Antônio de Pádua Braga*

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da UFMG - PPGEE UFMG, como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

UFMG - Belo Horizonte
Dezembro/2008

Agradecimentos

- À Deus que sempre providenciou tudo em minha vida. Em quem creio e confio plenamente, que me conduziu até aqui e me levará muito mais além.
- À minha esposa Cristiane a quem muito amo e que sempre me apoiou e me ajudou. Obrigado pela compreensão e paciência.
- Aos meus amados filhos que com sua doçura e inocência sempre encheram minha vida de alegria e entusiasmo, e sobretudo pelas vezes em que me tiraram dos estudos para brincar. Obrigado!
- Aos meus pais que sempre se preocuparam e se esforçaram para me dar uma educação de qualidade e ao meu irmão por todo o suporte que me dá. Amo vocês!
- Ao Professor Braga meu orientador, pela oportunidade, ajuda, paciência e entusiasmo no dia-a-dia.
- Aos amigos do LITC e CPDEE, por toda ajuda que me deram.
- Aos meus amigos e colegas da ESCHER Consultoria, em especial ao Jan pelos vários “debates matemáticos”.

Resumo

O câncer de mama é o tipo de câncer que mais atinge as mulheres no mundo e é o segundo tipo de câncer mais comum, perdendo apenas para o de pulmão. Existem diversos fatores que potencializam o risco de desenvolvê-lo e também que o minimizam como alimentação saudável e exercícios físicos. Contudo, a prevenção é mesmo a arma mais eficaz contra o câncer. Sua detecção prematura aumenta as chances do paciente se curar e passar por menos transtornos no tratamento.

O tratamento do câncer de mama constitui-se de três etapas, sendo o tratamento quimeoterápico neoadjuvante ou pré-cirúrgico, a cirurgia propriamente dita e o tratamento quimeoterápico adjuvante ou pós-cirúrgico. É fato que os tratamentos quimeoterápicos impõem grandes sofrimentos aos pacintes e no caso da quimeoterapia neoadjuvante cerca de 30% apenas dos pacientes apresentam alguma resposta positiva, seja na redução do tumor, seja na completa eliminação do mesmo dispensando a cirurgia.

O grande desejo dos médicos é dispor de algum meio para prever a resposta dos pacientes ao tratamento pré-cirúrgico, evitando assim, que aqueles que não responderiam bem ao tratamento precisem passar por todo este sofrimento desnecessariamente.

Os exames clínicos não são suficientes para se tomar a decisão de aplicar ou não o tratamento, mas com o desenvolvimento de novas técnicas de exames genéticos uma nova esperança surge e talvez seja possível desenvolver uma ferramenta que auxilie os médicos no tratamento de pacientes com câncer de mama.

O trabalho de Euler Horta [36] se dedica a selecionar as informações genéticas mais importantes dentre todas as obtidas pelos microarrays na previsão da eficiência da quimeoterapia neoadjuvante do câncer de mama. O trabalho que ora é apresentado trata do desenvolvimento e aplicação de um método semi-supervisionado aos conjuntos de dados (sondas) determinados no trabalho citado acima.

Foram coletados os dados de expressão genética de pacientes nos Estados Unidos e na França. Como o método implementado é semi-supervisionado parte dos dados serão considerados como rotulados e outra parte como não rotulados e a idéia é, a partir do treinamento de uma rede MLP com os dados rotulados, em conjunto com a informação de margem geométrica calculada tanto para os dados de treinamento como para os dados não rotulados, tentar chegar à uma solução que classifique bem ambos os conjuntos de dados.

A proposta deste trabalho é calcular geometricamente a margem dos dados de entrada para ser utilizado como informação de distribuição dos dados no método semi-supervisionado mas a grande dificuldade é calculá-la para o conjunto de dados não-rotulados. Para viabilizar o cálculo da margem geométrica foi necessário então desenvolver uma maneira de identificar os padrões limites entre as classes, utilizando o método de agrupamento Fuzzy (FCM) e o fatiamento do espaço dos padrões que é proposto neste trabalho.

A aplicação deste método semi-supervisionado batizado de *método semi-supervisionado baseado na margem geométrica (MSMG)* chegou a resultados muito bons e abre espaço para muitas abordagens e testes diferentes.

Abstract

Breast cancer is the most common cancer kind in women in all the world and it is the second most common kind of cancer, behind only of lung cancer. Exist many factors that can increase or reduce the risk of cancer like healthily alimentation and body exercises. However prevention is the most effective weapon against breast cancer. It's early detection increases patient cure chances and reduces exposition to damages of the treatment.

The breast cancer treatment has three stages, consisting in pre-operative chemotherapy, surgery and post-operative chemotherapy. As is known, chemotherapy imposes too much suffering to the patients, and, in the case of pre-operative chemotherapy, only 30% of them show some response to the treatment. It's desirable to have some way to preview the pre-operative chemotherapy response of the patients before they are subjected to it, avoiding unnecessary sufferings.

Clinical examinations are not enough to take the decision of subject the patient to the treatment or not, but nowadays new technics allows access to patient genetic information that can be used to try solve this problem and avoid them to suffers unnecessarily.

Euler Horta's work [36] separeted the most important genetic feature to the breast cancer pre-operative chemotherapy efficiency preview problem. Our work is a semi-supervised method applied to data set defined in cited work.

Patient genetic feature was collected in the USA and France. Some of this data set will be considered as labeled patterns and other as unlabeled patterns to be used in the semi-supervised method presented at this work. The idea is train one MLP network with the labeled data set, and use geometric margin calculated value for both labeled and unlabeled data set together, to select one solution that well classifies both data set.

This work proposal is evaluate margin geometrically to be the distribution feature to the semi-supervised method, however, the big difficult is on evaluate

it to the unlabeled data set. Was necessary develop some way to identify the patterns limits between class to evaluate margin geometrically. Was used the FCM method and the feature space slicing proposed in this work.

This method called *geometric margin based semi-supervised method (MSMG from portuguese método semi-supervisionado baseado na margem geométrica)*, had god results and inspires new approaches and tests.

Sumário

Lista de Abreviações

FCM	-	Algoritmo *Fuzzy C-means Method*
MSMG	-	Método semi-supervisionado baseado na margem geométrica
MLP	-	*Multi-Layer Perceptron*
MOBJ	-	Multi-objetivo
SMC-MOBJ	-	Método multi-objetivo de modos deslizantes
SVM	-	*Support vector machines*
MILFAT	-	Método de identificação de limites por fatiamento
MILP	-	Método de identificação de limites por probabilidade
MILFAT	-	Método de identificação de limites da classe *labeled*
PCR	-	*Pathologic complete response*
NOPCR	-	*No pathologic complete response*
RD	-	*Residual disease*
RNA	-	Redes neurais artificiais
VC	-	Dimensão Vapnik-Chervonenkis

Lista de Símbolos

$\mathbf{w}$ - Vetor de pesos da rede neural artificial
$\mathbf{e}$ - Vetor de erros da rede neural artificial
S - Superfícies deslizantes do algoritmo SMC-MOBJ
α - Ganho do algoritmo SMC-MOBJ
β - Ganho do algoritmo SMC-MOBJ
$\mathbf{X}$ - Vetor dos padrões de entrada
d - Saída esperada da rede neural artificial
y - Saída real da RNA
b - *Bias* da RNA
ρ - Margem de separação
ξ - Variáveis *Slacc* das máquina de vetores de suporte
P - Probabilidade de um padrão de entrada pertencer a uma partição do FCM
c - Número de partições definidas para o FCM
f - Número de faixas definidas para o MILFAT
k - Variáveis de folga na determinação dos padrões limites

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A principal característica em uma rede neural artificial é sua capacidade de aprender a partir do meio em que está inserida, e mais, de melhorar seu desempenho através de alguma forma de aprendizagem. Ao longo do tempo, uma rede neural, melhora seu desempenho de acordo com uma medida pré-definida em um processo interativo de ajuste de seus parâmetros. A cada iteração, em tese, a rede se torna mais "instruída". A definição de aprendizagem depende do ponto de vista de quem a está fazendo. No capítulo 2 é apresentada uma definição para aprendizado de máquina a partir da qual pode-se entender que a rede é "estimulada" pelas informações retiradas do ambiente, sofre modificações em seus parâmetros livres e por fim responde de maneira diferente e nova a este mesmo ambiente.

A forma com que a rede neural irá adquirir o conhecimento para solucionar o problema é constituída por um conjunto de regras e é chamado de algoritmo de aprendizagem. Existem diversos paradigmas de aprendizagem e o que interessa neste trabalho é o aprendizado semi-supervisionado. Este é aplicável sobretudo a uma classe de problemas de importância cada vez mais significativa no contexto da aprendizagem de máquinas que possuem um desbalanceamento entre o conjunto de dados de treinamento e o de teste, que pode ser função de um classificador muito oneroso ou devido à própria falta de informação. Desta forma não se pode garantir que as informações rotuladas representem completamente o sistema a ser aprendido. Isto restringe a utilização de algoritmos de aprendizagem indutivos (supervisionados), contudo tenta-se extrair o máximo de informação possível do conjunto de dados não rotulados na tentativa de melhorar a generalização da solução a ser obtida.

A aprendizagem semi-supervisionada conta então com um professor, ou

supervisor, que ao se apresentar determinados dados de entrada (informações do ambiente) à rede neural, informará qual a saída esperada e fará os ajustes necessários nos parâmetros da rede de forma a alcançá-la. Contudo, neste tipo de problema, existem dados para os quais o supervisor não saberá o que esperar na saída. Aqui então, o algorítmo deve tentar retirar alguma informação deste conjunto de dados (dos quais não se conhece a saída da rede) para ajustar os parâmetros da rede, de tal forma que os dados rotulados sejam classificados corretamente e aqueles não rotulados (de teste[1]) também sejam.

A proposta deste trabalho é justamente extrair uma informação consistente do conjunto de dados não rotulados de maneira que ela tenha uma aplicabilidade bem geral. Quando se trata de classificar corretamente os dados de entrada deseja-se definir uma separação entre uma classe de dados e outra por exemplo. Para cada problema de classificação vai existir uma gama de soluções que classificam corretamente as classes de dados, contudo, a melhor solução é aquela que está exatamente no meio da separação destas classes. A distância entre esta superfície de separação e os dados de uma classe é comumente chamada de margem e muitos métodos tentam definir a superfície de separação cuja margem seja máxima. Um destaque especial é dado, na literatura, para as SVMs, que identificam o hiperplano de separação de margem máxima.

Como está-se tratando de problemas que possuem dados não rotulados, estes podem não ser bem classificados pela solução de margem máxima obtida com o conjunto de dados rotulados apenas. Decidiu-se então fazer uma abordagem geométrica do problema da margem. O conjunto de dados possui dados rotulados, ou seja, padrões de entrada com uma respectiva saída esperada do sistema, e dados não rotulados, ou seja dados que não se sabe qual será a saída do sistema. Mas tem-se toda a distribuição espacial dos dados de entrada. O que se quer é definir áreas de baixa densidade de padrões de entrada, onde supõem-se estar a separação das classes. Desenvolveu-se então um método para tentar encontrar estas regiões de baixa densidade de padrões e, consequentemente, definir os possíveis limites das classes, para então calcular geometricamente a margem. Espera-se que a solução que apresente o valor máximo de margem geométrica, tanto para o conjunto de dados rotulados, como para o conjunto não rotulado, seja o hiperplano de separação ótimo ou próximo à ele.

Calcular geometricamente a margem de uma dada solução não depende do tipo de rede e nem de sua topologia. Pode ser feito tanto para SVMs quanto

[1]No capítulo 3 faz-se uma argumentação explicando como e porque é possível utilizar o conjunto de teste como transdutivo também.

para redes MLP. Isto mostra seu aspecto geral de aplicação. Atualmente a literatura enfatiza muito a utilização de SVMs quando se trata de classificadores de margem larga, contudo, este método semi-supervisionado pode ser utilizado com qualquer tipo de rede.

O método foi desenvolvido e é apresentado neste trabalho com dados de um problema do tipo *toy problem*, contudo, aplicá-lo em um problema real era de fundamental importância para validá-lo. Um conjunto de dados julgado interessante para aplicação deste método semi-supervisionado de aprendizagem foi o do projeto CAPES-COFECUB desenvolvido pelo Laboratório de Inteligência Computacional em conjunto com o pessoal da *Université Paris-Est* sobre o problema da previsão da eficiência do tratamento quimeoterápico neoajuvante em pacientes com câncer de mama. Os dados utilizados para desenvolver o método de aprendizagem baseado na margem geométrica eram bem comportados, de baixa dimensão e com uma região de separação muito bem definida, mesmo para os dados não rotulados, contudo, os dados de aplicação escolhidos são de alta dimensão e com uma separação que não é bem definida ou bem comportada.

Hoje o câncer de mama atinge mulheres no mundo inteiro, sendo este o tipo mais comum entre elas. Seu tratamento envolve três fases e todas elas igualmente duras e sofridas, seja no campo físico, seja no campo psicológico. Estas fases constituem-se primeiramente de um tratamento quimeoterápico pré-cirurgico (neoadjuvante), depois pela cirurgia e em seguida outro tratamento quimeoterápico (adjuvante).

Neste trabalho, os dados utilizados, referem-se aos problemas que envolvem a primeira fase: o tratamento neoadjuvante. Este tratamento tem básicamente um objetivo: iniciar o combate ao câncer de mama tentando reduzí-lo ou mesmo extinguí-lo completamente. Caso ele seja reduzido, não será necessário remover toda a mama, e se ele for extinto, a cirurgia não será necessária. Entretanto, nem todos os pacientes submetidos a este tratamento respondem de forma satisfatória. Aliás, apenas 30% destes apresentam boa resposta, seja reduzindo o tamanho do tumor ou eliminando-o.

Os demais 70% dos pacientes são submetidos ao tratamento e passam por todo o sofrimento e transtornos característicos inutilmente. Seria então muito interessante se ter uma maneira de prever, com segurança, se o paciente responderá bem ou não a esta primeira fase, para poder evitar que ele sofra desnecessariamente.

Exames clínicos não são suficientes para se tomar esta decisão, mas hoje, existe a tecnologia de microarrays que permitem levantar milhares de informações genéticas do paciente disponibilizando então uma grande massa de dados que podem, no futuro, viabilizar uma ferramenta de auxílio aos médi-

cos para a tomada desta decisão[2].

Os dados de aplicação deste trabalho foram obtidos de pacientes com câncer de mama nos Estados Unidos e na França. São 133 pacientes que tiveram cerca de 22 mil expressões de genes levantadas. Os trabalhos de Horta [36], Hess [35], Natowicz [45], Braga [44] [11], utilizam estes dados e têm dois objetivos principais: selecionar as sondas ou os genes que são mais relevantes para a solução deste problema de previsão (classificação em PCR e NOPCR, ou seja se o paciente responderá positivamente ao tratamento ou não) e aplicá-los a diversos classificadores tentando alcançar o melhor resultado possível tanto para o conjunto de dados destes pacientes definidos como conjunto de treinamento como para aqueles definidos como conjunto de validação.

1.1 Organização da dissertação

O capítulo 2 traz um pequeno resumo sobre os conhecimentos necessários ao entendimento do que será tratado ao longo do trabalho com várias referências a trabalhos que abordam os temas mais a fundo.

O capítulo 3 explica com detalhes o método semi-supervisionado baseado na margem geométrica a ser aplicado no problema.

O capítulo 4 mostra de onde vêm os dados e como eles foram selecionados e tratados.

O capítulo 5 apresenta os resultados obtidos com a aplicação do método semi-supervisionado proposto de diversas maneiras.

O capítulo 6 apresenta as conclusões do trabalho.

O apêndice A explica com mais detalhes o problema do câncer de mama e todas as suas implicações.

[2] O apêndice A explica um pouco melhor o problema de câncer de mama

CAPÍTULO 2

Revisão Bibliográfica

Neste capítulo pretende-se discorrer brevemente sobre os assuntos necessários para a compreensão do trabalho desenvolvido. São os conceitos básicos e as definições envolvidas, assim como a literatura utilizada para pesquisa e que também são fonte de explicações mais detalhadas.

2.1 O que são Redes Neurais Artificiais

Rede Neural artificial (RNA) é na verdade um modelo matemático do cérebro humano, sobretudo dos neurônios biológicos que o constituem. O cérebro é capaz de realizar tarefas extremamente complexas e não-lineares rapidamente e de forma paralela. Um exemplo bem claro disso é o processamento das informações fornecidas pela visão humana. O cérebro é capaz de identificar rapidamente em uma imagem adquirida pelos olhos tudo o que ele **conhece** e além disso é capaz de reagir e interagir com o ambiente a partir desta informação. Este **conhecimento** é adquirido através de **experiências**. O cérebro humano é capaz de realizar estas tarefas pois passou por experiências ao longo da vida e registrou o conhecimento e as regras de comportamento modelando a maneira como ele deve interpretar e reagir aos estímulos.

Em [34] uma rede neural artificial é definida como sendo um processador paralelo formado por unidades simples capazes de armazenar conhecimento adquirido através de experiências e de torná-lo disponível para uso. O conhecimento é adquirido e armazenado, assim como no cérebro humano, através de um treinamento ou processo de aprendizado que atua nos pesos das conexões (ou forças de conexão) entre suas unidades. Estes pesos são também conheci-

dos como pesos sinápticos. Ou seja, as redes neurais possuem a capacidade de aprender por exemplos e por conseqüência fazer extrapolações e interpolações. Esse aprendizado dito conexionista tenta determinar a intensidade das conexões entre eles, definindo valores dos pesos de cada uma destas conexões [12]. Para tanto, é necessário uma forma de treinar estas redes. Existe uma grande quantidade de algoritmos de aprendizado e cada um tem suas vantagens e desvantagens.

Maiores detalhes sobre os modelos de neurônios e arquiteturas de redes neurais podem ser encontrados em [34] e [12].

2.2 Aprendizado de máquina

Uma RNA, para ser utilizada, deve necessariamente passar por uma fase de treinamento, que é quando ela vai extrair todas as informações mais relevantes dos dados apresentados à rede, levando-a a criar uma representação própria do problema. Com isso ela será capaz não somente de repetir algum comportamento visto durante o treinamento, mas de tomar decisões acerca de situações que ela não havia tomado contato antes. Esta é uma característica importante das RNAs: elas são capazes de se adaptar a novos ambientes e de resolver novos problemas.

Existem diversas maneiras de se caracterizar o aprendizado de máquina, contudo o mais importante talvez seja o aprendizado indutivo. *O processo de indução é a forma de inferência lógica caracterizado pelas conclusões generalistas obtidas através de exemplos particulares* [41].

Mas o que vem a ser aprendizado de máquina?

Aprendizado é um processo pelo qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual a rede está inserida. O tipo de aprendizado é determinado pela maneira pela qual a modificação dos parâmetros ocorre [42].

Aprendizagem de uma rede neural se dá através de um processo interativo de ajustes aplicados aos seus pesos sinapticos e níveis de bias[34]. São estes parâmetros (os pesos das conexões) que armazenam o conhecimento que a rede adquire no processo. Existem diversos métodos para treinamento de redes que podem ser agrupados em três linhas distintas :

- Aprendizado supervisionado;
- aprendizado não-supervisionado e;
- aprendizado semi-supervisionado.

2.2.1 O Aprendizado Supervisionado

O aprendizado supervisionado é o mais comum empregado em treinamento de redes neurais. Recebe esse nome justamente por receber os dados com os valores de entrada e saída desejadas para a rede através de um supervisor externo. O algoritmo deve tentar então ajustar os pesos da rede de forma a conseguir produzir uma saída igual à desejada para o dado de entrada em questão. A apresentação de cada padrão de entrada à rede gera uma resposta, que é comparada à resposta desejada pelo supervisor. Caso haja diferença entre estas saídas, o supervisor ajustará os pesos das conexões da rede na tentativa de minimizar este erro. Esses ajustes são feitos de forma incremental. A figura 2.1 mostra o diagrama em blocos da aprendizagem com supervisor.

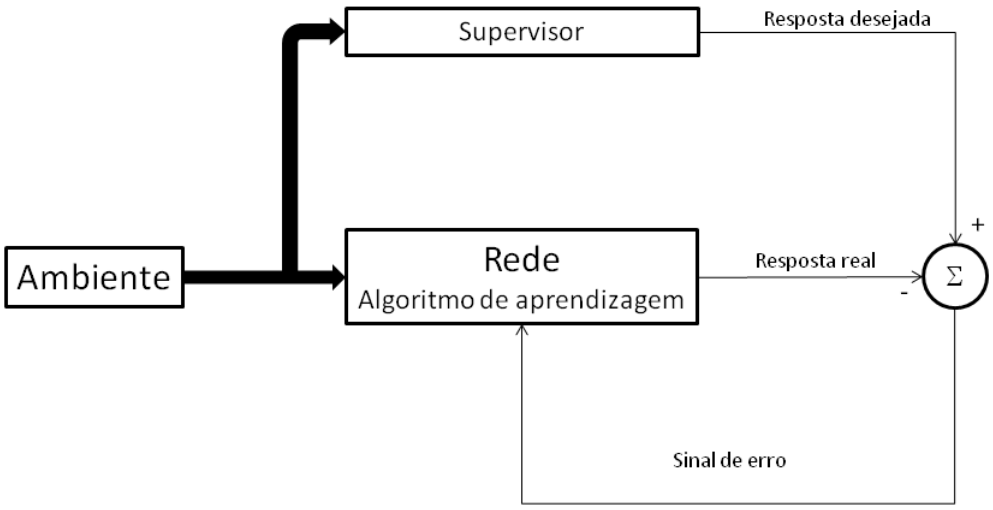


Figura 2.1: Diagrama de blocos da aprendizagem supervisionada

Este tipo de aprendizado pode ser implementado de duas formas: off-line e on-line. Como o próprio nome diz, o treinamento off-line é realizado com um conjunto de dados fixo que não se altera, enquanto que no treinamento on-line a rede é treinada com dados que podem ser alterados e acrescentados durante o processo. Uma desvantagem do aprendizado supervisionado é que na ausência do supervisor a rede não conseguirá aprender novas estratégias para situações não contempladas no treinamento.

2.2.2 O Aprendizado Não-Supervisionado

Neste tipo de aprendizado não existe a figura do supervisor que o orienta. Ao contrário, a rede tem que tentar aprender somente com a informação dos padrões de entrada. Ela não dispõe da informação de saída desejada para poder comparar com a saída real e ajustar seus pesos sinápticos. Este aprendizado só é possível com a presença de redundância nos dados de entrada. Este processo de aprendizado é também conhecido como clustering que será dicutido com mais detalhes no ítem 2.5. Da mesma forma que no aprendizado supervisionado, existem diversas topologias e algoritmos para realizar este treinamento.

2.2.3 O Aprendizado Semi-Supervisionado

No aprendizado semi-supervisionado dispõe-se de dados de entrada rotulados e de dados não rotulados, ou seja, existe um conjunto de dados, que irão alimentar as sinapses da rede, para os quais se conhece o valor de saída esperado. Para esses dados existirá a figura do supervisor que comparará a saída real da rede com o valor de saída esperado (rótulos) e utilizara o erro dessa medida para ajustar de forma incremental os pesos das conexões das redes. Existe também um outro conjunto de dados de entrada que faz parte da caracterização do problema, para os quais não se conhece o valor de saída da rede. Para esses dados a figura do supervisor é inútil. Os métodos semi-supervisionados são aqueles que supervisionam os dados rotulados e que tentam extrair informações dos conjuntos de dados que não se conhecem as classificações para auxiliar no ajuste das redes.

2.3 Redes MLP

Redes de perceptrons de múltiplas camadas (do termo *multi layer perceptron* em Inglês) têm a característica de solucionar problemas não linearmente separáveis, pois utilizam uma ou mais camadas intermediarias ou escondidas. Contudo simplesmente implementar o treinamento de uma função não implica em garantia de generalização da mesma, pois a rede pode convergir para o mínimo local. Existem diversos algoritmos e estratégias de treinamento para essas redes [50],[25],[48],[49],[47],[46],[33], a maioria dos algoritmos são supervisionados e podem ser classificados em dois tipos, estáticos e dinâmicos. Os estáticos ajustam valores dos pesos, mas não alteram a estrutura da rede, enquanto que os dinâmicos podem fazer as duas coisas. O algoritmo mais conhecido desse tipo de rede é o back-propagation [50], e dele derivam a maioria dos métodos de aprendizagem dessas redes.

2.4 MOBJ

O problema da otimização multi-objetiva consiste *na obtenção de um conjunto viável de soluções que satisfaça algumas restrições e otimize uma função vetorial, constituída por diversos termos ou funções objetivo escalares* [40]. Em outras palavras, existem diversos problemas onde é necessário encontrar soluções que precisam ser balanceadas quanto há diversos aspéctos que implicam em vários critérios e restrições. Por exemplo, em muitos casos, pode-se treinar uma rede MLP preocupando-se apenas em minimizar o erro quadrático, ou seja, a diferença entre a saída da rede e o valor desejado desta, sem se preocupar com a "dimensão" ou a sua complexidade. Mas talvez seja mais interessante uma rede que tenha complexidade menor e com erro baixo porém um pouco maior que o erro de uma rede de alta complexidade. Assim, em muitos casos, quer ou precisa-se "firmar" um compromisso entre erro e complexidade (norma). Desta maneira diversas outras funções e restrições podem ser contempladas em uma abordagem multi-objetiva.

Um problema multi-objetivo apresenta diversas funções objetivos que devem ser otimizadas e uma série de restrições que devem ser respeitadas. A formulação matemática para este tipo de problema pode ser encontrada com detalhes em [51], [26] e [38].

2.4.1 Pareto ótimo

Vários fatores e critérios influenciam na solução de um problema multi-objetivo. Resolver um problema destes implica, em algum momento, tomar uma decisão. Para se entender melhor o conceito de pareto ótimo lança-se mão de um exemplo simples de problema multi-objetivo. Seja a seguinte questão: alguém deseja comprar uma televisão LCD com o menor tempo de resposta possível para garantir que a imagem não terá os indesejáveis "fantasmas" e, é claro, com o menor preço possível. Após uma pesquisa de mercado chega-se aos seguintes modelos e preços de TVs mostrados na figura 2.2.

Resolver este problema então implica em ter que tomar uma decisão. Aquele que primar pela qualidade da imagem com um menor tempo de resposta tem que estar disposto a pagar mais caro por isso, enquanto que outro com sérias restrições orçamentárias tem que se contentar com uma qualidade inferior de imagem. Uma solução balanceada, nem tão cara e nem tão ruim em termos da qualidade da imagem seria escolher a TV B, de R$ 3.000 que tem 3ms de tempo de resposta. Contudo, este gráfico também mostram duas opções que não teriam sentido serem escolhidas nesta conjuntura: as TVs C e D. A TV D tem um tempo de resposta muito pior que o da TV B pelo mesmo preço praticamente, e a TV C tem um tempo de resposta igualmente pior mas é mais

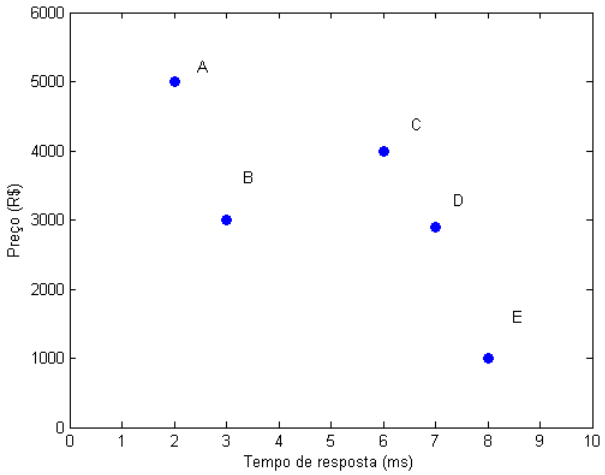


Figura 2.2: Exemplo de custo x benefício de TVs

cara que a TV B. Pode-se dizer então que as soluções C e D são dominadas pelas demais soluções.

Introduz-se assim o conceito de dominância do pareto, ou seja, soluções em que todos objetivos são claramente melhores que os de outras soluções dominam estas últimas. O pareto é então formado por soluções que não são dominadas por nenhuma outra e que dominam as demais. As soluções do pareto não são melhores umas que as outras, mas apresentam um "compromisso" entre elas. No exemplo da figura 2.2 o pareto é formado pelas soluções A, B e E e não pode-se afirmar que uma é melhor que a outra, apenas que enquanto uma tem melhor qualidade e maior preço, a outra tem menor qualidade mas preço menor.

2.4.2 SMC-MOBJ

O que se deseja de uma rede neural treinada é que haja um balanço entre os erros de classificação cometidos no conjunto de treinamento e no conjunto de teste, e claro, que estes erros sejam mínimos.

O algoritmo Multi-Objetivo [16] tenta alcançar um equilíbrio para o dilema *bias and variance* de uma rede neural, selecionando soluções no Pareto no espaço dos objetivos norma do vetor de pesos da rede $||\mathbf{w}||$ e erro do conjunto de treinamento $\mathbf{e}$. Contudo existe um algoritmo de modos deslizantes capaz de gerar trajtórias arbitrárias no espaço dos objetivos que permite alcançar um ponto qualquer $(\mathbf{e}_t, ||\mathbf{w}_t||)$ no pareto, minimizando a distância entre a solução atual da rede em uma determinada iteração e o ponto desejado.

Este algoritmo pode alcançar qualquer solução $(\mathbf{e}_k, \|\mathbf{w}_k\|)$ no espaço de objetivos definido pelo somatório do erro quadrático $\mathbf{e}_k$ e pela norma do vetor de pesos $\|\mathbf{w}_k\|$. Pode-se portanto, gerar qualquer trajetória no espaço dos objetivos

com o algoritmo SMC-MOBJ. Este algoritmo minimiza duas superfícies de modos deslizantes. Estas superfícies deslizantes são definidas por $S_v = (\mathbf{e} - \mathbf{e}_k)$ e por $S_{\|\mathbf{w}\|} = \left(\|\mathbf{w}\|^2 - \|\mathbf{w}_k\|^2\right)$ e, em conjunto com os ganhos α e β controlam a convergência dos pesos da rede. O que é feito é simplesmente o ajuste da direção dos gradientes locais, de acordo com a posição atual e aquela onde se quer chegar no espaço dos objetivos. Maiores informações e toda formulação sobre o SMC-MOBJ podem ser obtidas em [16].

2.4.3 SVM - um classificador de margem larga

As máquinas de vetores de suporte são uma categoria de rede alimentadas adiante universais, que podem ser também utilizadas para a classificação de padrões e regressão linear. Elas foram propostas por Vapnik [63], [64], [66] e [67].

SVM mapeia através de um produto interno (Kernel) os dados de entrada para um espaço característico de alta-dimensão, onde mesmo que a entrada seja não-linear, a separação pode ficar linear. Um hiperplano ótimo é construído para separar os dados em duas classes. Quando os dados de treinamento são separáveis, o hiperplano ótimo no espaço característico apresenta uma máxima margem de separação. Caso apresentem superposição (dados não separáveis), uma generalização deste conceito é utilizada.

A SVM é baseada nos princípios da minimização do risco estrutural, que tem suas origens na teoria do aprendizado estatístico. É considerado que o erro do algoritmo de aprendizagem junto aos dados de validação (erro de generalização), é limitado pelo erro de treinamento mais um termo que depende da dimensão VC (dimensão Vapnik e Chervonenkis) [54], que é uma medida da capacidade de expressão de uma família de funções. O que se quer é construir um conjunto de hiperplanos tendo como estratégia a variação da dimensão VC, de modo que o risco empírico (erro de treinamento) e a dimensão VC sejam minimizados ao mesmo tempo.

Uma SVM é treinada por um algoritmo de otimização quadrático que garante a convergência para um mínimo global da superfície de erro (diferença entre a saída da rede e a saída desejada). Transforma-se o problema de otimização primal em sua representação dual permitindo que o problema de dimensionalidade não seja mais uma dificuldade. Assim, o número de parâmetros ajustados não dependerá mais do número de atributos sendo utilizados, ou seja, da dimensão do espaço a que pertencem os dados de treinamento.

Mais detalhes sobre as SVMS, sua formulação, suas aplicações e sobre margem podem ser encontrados em [54], [13], [64], [14], [39], [60], [61], [43], [31] e [19].

O artigo [3] tenta dar uma explicação intuitiva para as SVMs a partir de

uma perspectiva geométrica. Ele explica os conceitos de margem que parecem com o que se propõe neste trabalho. Mas aqui ele usa um conceito de limites (envoltória) do conjunto convexo dos dados e do conjunto reduzido destes dados para o caso não linear. Ele encontra a envoltória dos dados do conjunto convexo ou do reduzido e calcula a reta que é a bisetriz da reta que liga os pontos mais próximos um do outro conjunto direto sem ficar avaliando a margem de cada solução.

Um outro artigo encontrado na literatura que pode ser interessante para o leitor é o de Bennett de título *Semi-supervised suport vector machines* [5]. Ele propõem uma SVM semi-supervisionada.

Um exemplo de como utilizar as SVMs para clustering é demonstrado em [2]. Já em [9] tem-se uma demonstração de como utilizar uma SVM para regressão.

2.4.4 O hiperplano ótimo

Para facilitar a compreensão do que vem a ser o hiperplano ótimo de separação tratar-se-á inicialmente do caso de padrões de entrada linearmente separáveis, pois assim pode-se explicar a idéia básica que rege uma SVM em um cenário simples.

Seja um conjunto de dados de treinamento $(\mathbf{X}_i, d_i)_{i=1}^{N}$ onde $\mathbf{X}_i$ é o i-ésimo padrão de entrada, d_i é a saída desejada da rede para este padrão de entrada [34] e N é o número total de padrões de entrada. A equação de uma superfície de separação entre duas classes $d_i = -1$ e $d_i = +1$ linearmente separáveis é dada por 2.1 onde $\mathbf{w}$ é o vetor de pesos da rede e b é o bias.

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{2.1}$$

Daí pode-se reescrever a equação 2.1 nas equações 2.2.

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b \geq 0 \; para \; d_i = +1 \\ \mathbf{w}^T\mathbf{x} + b < 0 \; para \; d_i = -1 \end{aligned} \tag{2.2}$$

A margem de separação ρ é definida como sendo a distância entre o hiperplano de separação (determinado por $\mathbf{w}$ e b) e o dado (padrão de entrada) mais próximo. O objetivo de uma máquina de vetores de suporte é encontrar a separação ótima, ou seja, determinar o vetor $\mathbf{w}$ de pesos da rede e o bias b de forma que ρ seja máximo. A figura 2.3 mostra o que vem a ser o hyperplano ótimo, os vetores de suporte e o conceito de margem.

A questão é então encontrar $\mathbf{w}$ e b que determinem o hiperplano ótimo. Caso

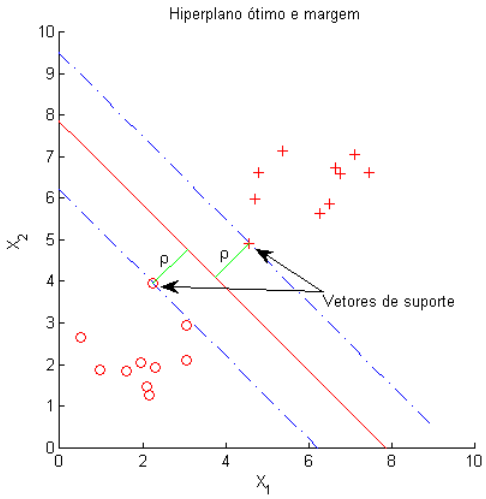


Figura 2.3: Exemplo de um hiperplano ótimo para padrões linearmente separáveis

os padrões sejam linearmente separáveis, como suposto neste ítem, pode-se sempre reescalar $\mathbf{w}$ e b de maneira que a equação 2.2 se apresente como[1] na equação 2.3 [34].

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b \geq 1 \; para \; d_i = +1 \\ \mathbf{w}^T\mathbf{x} + b \leq -1 \; para \; d_i = -1 \end{aligned} \tag{2.3}$$

Os padrões que serão chamados vetores de suporte são aqueles padrões de entrada $(\mathbf{X}_i, d_i)$ para os quais a primeira ou segunda linha da equação 2.3 for satisfeita com a igualdade. Conceitualmente eles são aqueles pontos mais próximos ao plano de separação e por isso mesmo são os mais difíceis de se classificar, pois têm os maiores erros quadráticos. São eles também que acabam representando todo o conjunto de padrões de entrada, ou seja, treinar a rede com todos os padrões, ou só com eles leva à mesma superfície de separação ótima.

Através de um desenvolvimento algébrico [34] pode-se chegar à equação 2.4.

$$\rho = \frac{2}{\|\mathbf{w}\|} \tag{2.4}$$

A equação 2.4 mostra que maximizar a margem ρ significa minimizar a norma do vetor de pesos $\mathbf{w}$. Disso tudo conclui-se que o hiperplano ótimo, de margem funcional máxima, é obtido ao se minimizar a norma do vetor de

[1] Bishop, em seu livro [10], dá uma explicação mais clara para a formulação do hiperplano ótimo e conseqüentemente da margem, preenchendo algumas lacunas que aparecem em outras fontes como [34], [55] e [58].

pesos da rede.

E como fica a margem para o caso do conjunto de dados de entrada não ser linearmente separável? Neste caso não é possível construir um hiperplano de separação sem incorrer em erros de classificação. O objetivo então é construir um que minimize estes erros e maximize a margem.

Para se resolver este problema é necessário introduzir, na definição do hiperplano de separação, um conjunto de variáveis escalares não negativas $\{\xi\}_{i=1}^{N}$, onde N é o número de padrões de entrada. Assim, a fórmula que define o hiperplano de separação fica como na equação 2.5.

$$d_i \left(\mathbf{w}^T x + b\right) \geq 1 - \xi_i \tag{2.5}$$

Estas variáveis, também conhecidas como *slacc variables*, medem o grau de desvio de um padrão de entrada do hiperplano de separação ótimo. Para $0 \leq \xi \leq 1$ o padrão em questão está dentro da faixa da margem , mas do lado correto da separação, e se $\xi > 1$ então ele está classificado incorretamente. Se estes padrões forem deixados de fora do treinamento, os vetores de suporte não mudarão. Isto mostra que os vetores de suporte são definido da mesma maneira, seja para o caso linearmente separável ou não.

A formulação e a implementação destes cálculos podem ser encontrados em [34].

Existem diversos trabalhos e estudos sobre a margem na literatura. Cita-se alguns trabalhos encontrados que abordam parcial ou totalmente este assunto:

- O capítulo 1 do livro de Dale [21] se destina a dar uma introdução aos classificadores de margem larga e também boas noções sobre SVMs.
- Os livros de Haykin [34], Smola [58] e Bishop [10] também abordam este assunto.
- O artigo de Bennett [4] foca a regressão de dados, porém, em sua introdução existem definições úteis. Aqui, a autora define como sendo a solução de margem máxima, aquela que gera um plano bissetriz à reta que une os pontos mais próximos dos limites de cada classe, para o caso de dados linearmente separáveis, e usa a distância entre os pontos mais próximos dos limites reduzidos para o caso de dados não linearmente separáveis.
- Outro artigo de Bennett [3] tenta dar uma explicação mais intuitiva sobre SVMs apartir de uma perspectiva geométrica.
- Em [29] pode-se encontrar um classificador de margem larga utilizando perceptrons. É interessante pois seu classificador é simples de imple-

mentar e parece funcionar bem (pelo menos para o caso linearmente separável).

- Já em [30] propõe-se um novo algoritmo incremental de classificação que aproxima às soluções das SVMs, sendo um pouco inferiores.

- Outros artigos que abordam esta questão são os [39], [69], [52], [37], [28], [53], [15], [64], [63], [43], [7], [70], [19], [56], [59], [6], [62].

2.5 Clustering

O paradigma do aprendizado não-supervisionado, ou clustering, consiste em determinar ou identificar conjuntos de dados que guardem uma certa similaridade, ou seja, aqueles padrões de entrada que são mais próximos, e que de alguma forma serão agrupados. Desta maneira, padrões que, ao final, fazem parte de um mesmo grupo (cluster) são os mais similares entre si de acordo com a medida de distância (critério de similaridade) adotado. Obviamente estes mesmos padrões guardam menos similaridade com padrões que pertençam a outros agrupamentos.

Jain define em [27] quais os principais passos na solução de um problema de clustering que são listados a seguir:

1. Representação dos padrões

   Nesta etapa está-se preocupado em organizar e extrair as características dos dados de entrada de maneira que sejam apresentadas aos processos de agrupamento (clustering) apenas aquelas mais importantes e relevantes destes dados. Aqui se determina o número de classes a se considerar, o número de dados disponíveis e o número, tipo e escala das características destes dados a serem informados ao algoritmo.

2. Definição de uma medida de similaridade

   Já nesta etapa o que é definido é a medida de similaridade que será adotada para o agrupamento. Esta medida de similaridade é mesmo uma função de distância entre os pares de dados de entrada. Existem diversas medidas de distância que podem ser utilizadas, todas com suas vantagens e desvantagens e obviamente cada uma se aplica melhor a um determinado tipo de problema[2].

3. Agrupamento (clustering)

[2] Em [27] encontra-se uma lista destas distâncias.

Nesta fase é que se faz o agrupamento própriamente dito. Existem diversas técnicas de agrupamento e vários aspéctos a serem considerados. Pode-se dividir os algoritmos em dois grupos: os hierárquicos e os não-hierárquicos (particionais). Eles podem ser:

(a) Associativos ou desassociativos

O primero parte dos padrões separados e vai agrupando e o segundo parte dos padrões agrupados e vai dividindo;

(b) Monothetic ou polythetic

O primeiro utiliza uma característica por vez dos padrões para calcular a distânca, enquanto que o segundo utiliza todas as informações e características de cada padrão para determinar a distância ao mesmo tempo.

(c) Discreto ou difusso

O primeiro associa cada padrão a uma única classe enquanto que o segundo associa um padrão a uma classe com um determinado grau de pertinência.

(d) Determinístico ou estocástico

(e) Incremental ou não-incremental

O primeiro lida com dados muito grandes e o segundo não.

Quanto aos algoritmos, também se encontram diversos na literatura como algoritmos de erro quadrático tais como o K-médias e o Graph-teoretic clustering, o Mixture-resolving e Mode-Seeking, o Nearest Neighbor Clustering, o Fuzzy Clustering, e outros. Em [27] pode-se encontrar uma pequena explicação sobre estes algoritmos.

4. Abstração (quando necessário)

A abstração é o ato de extrair uma representação simples e compacta do resultado obtido. Ela é importante pois facilita a compreensão humana, ajuda na compactação de dados e aumenta a eficiência nas decisões.

5. Validação dos resultados

Como o próprio nome do ítem diz, nesta etapa se faz a verificação e validação dos resultados de saída se necessário.

O material produzido por Jain [27] é uma boa fonte de informações sobre tudo o que envolve clustering. Outras fontes interessantes são [10], [24], [69], [23], [2], [18], [5], [22], [68], [62].

Neste trabalho, utiliza-se um método de agrupamento chamado FCM do termo em inglês *Fuzzy C-means*. Originalmente implementado por Dunn, seu

algoritmo transforma a função objetivo de erro quadrático em uma função objetivo fuzzy a ser minimizada. Bezdec [8] em 1981 generalizou esta função objetivo fuzzy introduzindo um expoente ponderador na mesma. Este método permite que cada padrão pertença a mais de um cluster com um certo grau de pertinência.

O FCM usa uma otimização interativa da função objetivo baseada na medida ponderada de similaridade entre o dado de entrada e o centro dos *clusters* definidos ou calculados. Inicialmente, dado um conjunto de dados a ser agrupado, seleciona-se o número de *clusters* que se quer encontrar, depois dá-se uma condição inicial, então, para cada iteração, calcula-se o centro dos *clusters* e atualiza-se a matriz de probabilidades de cada padrão pertencer ou não a um determinado *cluster*. Este algoritmo vai convergir para um mínimo local ou um *saddle point* da função objetivo. Em [8] pode-se encontrar toda a formulação deste método e uma explicação mais detalhada.

2.6 Considerações finais

Neste capítulo deu-se uma visão geral sobre os assuntos, definições e conceitos básicos que são relevantes ao entendimento do trabalho desenvolvido e que será explicado nos próximos capítulos. Aconselha-se fortemente a consulta aos materiais citados ao longo do texto para um maior aprofundamento.

CAPÍTULO

3

Descrição da Metodologia

Nos capítulos anteriores foi apresentado o problema e os objetivos deste trabalho. O método semi-supervisionado baseado na margem geométrica (MSMG) que foi implementado neste trabalho,será explicado neste capítulo e foi aplicado ao problema de previsão da eficiência da quimioterapia neoajuvante no tratamento de câncer de mama. Os resultados desta aplicação serão apresentados no capítulo seguinte. Contudo, para facilitar a explicação, a visualização e conseqüentemente o entendimento do método, neste capítulo, será utilizado um problema bem conhecido, de apenas duas dimensões (duas variáveis de entrada): o problema das duas luas. A figura (3.1) mostra os dados de entrada do problema exemplo.

Os setenta e nove padrões de entrada representados em azul na figura 3.1 serão considerados como sendo os dados do conjunto de treinamento e/ou

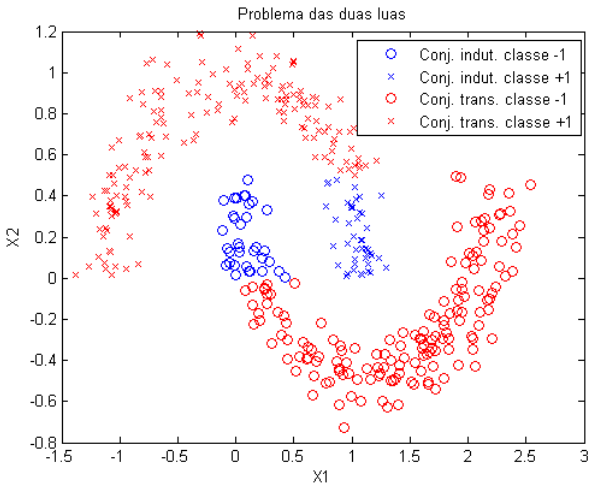


Figura 3.1: Problema das duas luas

indutivo. Já os trezentos e vinte e um dados de entrada representados por uma bolinha vermelha na figura serão considerados como sendo os dados do conjunto de validação e/ou transdutivo. Tem-se um problema a se resolver. O que se quer fazer é treinar de alguma forma uma rede, de tal forma que ela consiga, ao final, classificar da melhor forma possível, os padrões de entrada. Mas acontece que tem-se um conjunto de dados (indutivo) que se conhece sua classificação (rótulos), e, um outro conjunto (transdutivo) que não se conhece os rótulos. Com os dados do conjunto indutivo pode-se treinar a rede minimizando o erro quadrático entre a saída da rede e o rótulo conhecido, como também pode-se criar uma outra função de custo qualquer que se queira minimizar.

Já para o conjunto de dados transdutivo não se conhece sua classificação o que impede a minimização do erro. Contudo pode-se tentar utilizar alguma informação da distribuição destes dados que auxilie na escolha da melhor solução dentre as encontradas com o conjunto indutivo. Além disso pode-se também extrair esta mesma informação adicional do conjunto de treinamento, na tentativa de se ter um decisor ainda melhor.

Para este trabalho, a informação que se tentará calcular é a margem geométrica, de maneira que seu valor máximo indique uma boa solução. Ter uma solução de margem máxima significa que tem-se um hiperplano de separação passando bem no centro da região entre as classes que se quer separar. A seção 3.1 explica com mais detalhes o que vem a ser a margem.

Uma solução que atenda ao conjunto transdutivo (solução de margem geométrica máxima) necessariamente tem que atender aos critérios de margem máxima e erro mínimo do conjunto indutivo. Para o conjunto de dados indutivos, ao treinar as redes, será informado a classificação que temos a priore (classe 1 ou -1) enquanto que para o conjunto de dados transdutivos, não será considerado a informação de classificação que se conhece de antemão, porém ao validar a rede ela será utilizada.

Pode soar estranho utilizar o mesmo conjunto de dados para se extrair informações sobre a distribuição dos padrões que, aliada às informações de distribuição e à minimização do erro do conjunto indutivo permitirá selecionar a melhor solução para o problema e depois reutilizá-lo para a validação, contudo não há problemas nisto. Existe um conjunto de dados onde se conhece seus rótulos e onde é possível então treinar a rede e minimizar o erro. Além disso pode-se calcular a margem geométrica deste conjunto. Por outro lado, existe um conjunto de dados que, em tese, não se conhece sua classificação e de onde se pode tentar identificar as regiões de baixa densidade em relação ao número de padrões, para tentar forçar a solução a “passar” por esta região. Portanto, não se utiliza a informação da classificação deste conjunto que ex-

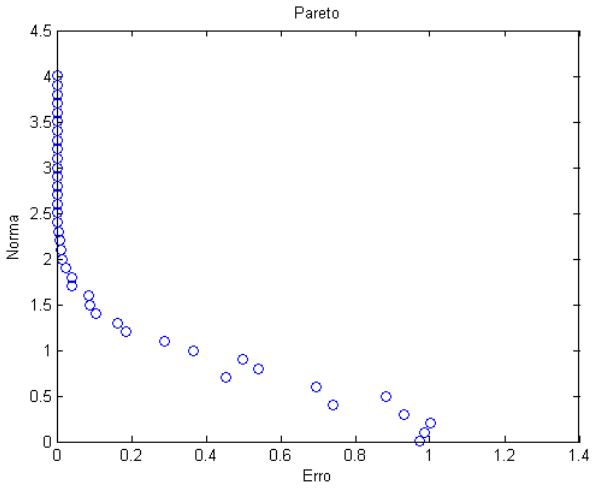


Figura 3.2: Pareto do problema das duas luas

iste previamente, o que permite utilizar este mesmo conjunto para validar o resultado e verificar se o conjunto transdutivo foi bem classificado.

O primeiro passo então é gerar o pareto com as soluções obtidas por um método multi-objetivo que vai minimizar o erro quadrático e a norma da rede. Na figura 3.2 pode-se observar o pareto gerado para este problema.

Soluções de maior norma e menor erro tendem a classificar muito bem todos os padrões de entrada do conjunto indutivo, mas tendem a não generalizar bem. Já as soluções de menor norma e maior erro quadrático tendem a errar mais na classificação dos padrões indutivos e a generalizar melhor quando aplicadas ao conjunto transdutivo (validação). Para o problema de exemplo em questão (duas luas) considerou-se uma rede MLP com apenas uma camada escondida de três neurônios. A camada escondida neste problema tem como principal função mapear os padrões de entrada em um espaço dimensional superior, que, se for de grau suficientemente maior poderá apresentar uma superficie de separação linear segundo o teorema de Cover [17].

É compreensível imaginar que a melhor solução represente um compromisso da norma com o erro quadrático incorrido. Mas pode-se imaginar também a seguinte questão: será que se pode extrair alguma informação do conjunto transdutivo que auxilie na decisão de qual solução é a melhor? Claro que parte-se do pressuposto de que não se conhece os rótulos destes dados. Ou indo um pouco mais além, será que esta característica pode ser considerada na função dos objetivos do método multi-objetivo, que, ao ser minimizada, consiga encontrar soluções melhores?

A idéia é procurar determinar os limites das regiões de baixa densidade de padrões e calcular a margem geométrica do conjunto, para tentar maximizá-la. Procura-se determinar os limites da região de baixa densidade de padrões no espaço de entradas, pois elas podem indicar que ali há uma transição de

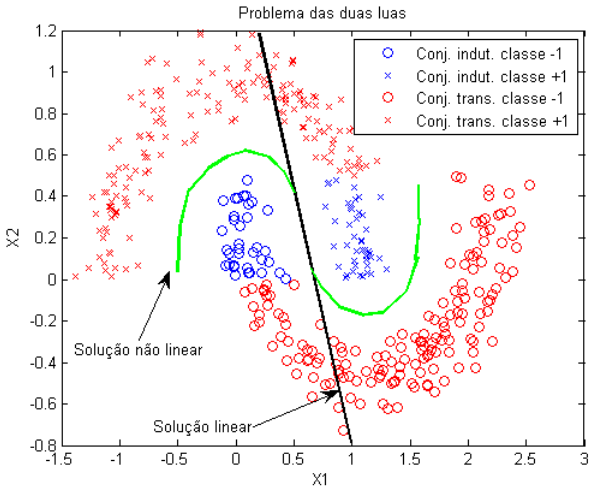


Figura 3.3: Exemplo de soluções do conjunto indutivo e transdutivo

classes, ou seja, espera-se que o hiperplano ótimo de separação esteja exatamente nestas zonas de baixa densidade uma vez que o padrão que pertence a uma determinada classe tende a estar mais afastado das demais classes. É claro que dentro de um conjunto de dados de uma mesma classe podem haver regiões de baixa densidade o que pode introduzir um erro, mas o método proposto se mostra robusto a esta questão e mesmo assim apresenta bons resultados.

Retornando ao problema das duas luas, é fato que as soluções do pareto geradas pelo método multi-objetivo devem ser, para o conjunto de dados de treinamento em questão, retas no espaço de entrada. Espera-se que, partindo de uma dada solução do pareto, e "caminhando" paralelamente ao eixo da norma, em direção a normas de valores maiores, exista alguma solução que tenha margem geométrica máxima no conjunto de dados indutivos e transdutivos e de mesmo erro calculado com o conjunto indutivo. Na figura 3.3 pode-se entender melhor o que foi explicado. Observa-se que o erro quadrático do conjunto de dados indutivo para a solução linear e a não-linear da figura é praticamente o mesmo, contudo a solução linear não generaliza bem.

Por esta razão foi preciso gerar um grid de soluções do conjunto de dados indutivo, para que se pudesse avaliar o método semi-supervisionado. Ele calculará a margem geométrica dos conjuntos de dados transdutivo e indutivo, para cada solução do grid, e verificar-se-á se a solução de margem geométrica máxima corresponde à melhor solução ou ao menos a uma solução plausível (validação).

Para gerar este grid de soluções usou-se um método para construção de soluções viáveis no espaço dos objetivos por meio do algoritmo multi-objetivo de controle de modos deslizantes [16]. Como descrito em 2.4.2 este algoritmo pode alcançar qualquer solução $(\mathbf{e}_k, \|\mathbf{w}_k\|)$ no espaço de objetivos definido pelo

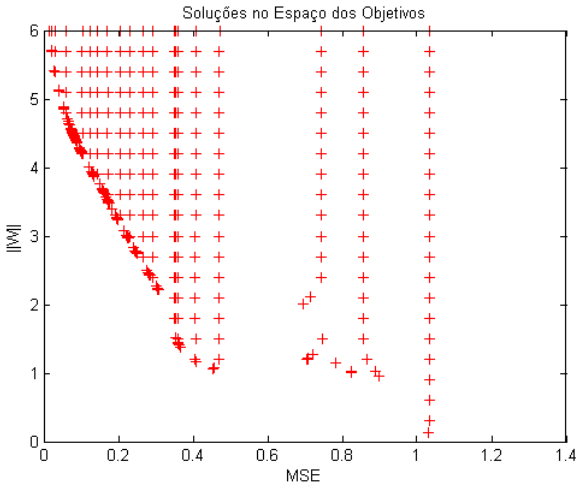


Figura 3.4: Grid de soluções no Espaço dos Objetivos

somatório do erro quadrático $\mathbf{e}_k$ e pela norma do vetor de pesos $\|\mathbf{w}_k\|$. Pode-se portanto, gerar qualquer trajetória no espaço dos objetivos com o algoritmo SMC-MOBJ. A idéia é então definir pontos alvo para o SMC-MOBJ de maneira que, ao final, tenha-se um grid de soluções por todo espaço dos objetivos, e então poder-se especular sobre qual é a melhor solução, ou seja, qual é a de margem máxima e se ela atende bem aos conjuntos de treinamento e de teste.

Partindo então das soluções encontradas no pareto montou-se um grid de normas do vetor de pesos e de erros quadráticos afim de que o algoritmo de modos deslizantes tente encontrar as soluções que atendam a estes critérios, ou seja, para um determinado erro quadrático $\mathbf{e}_k$ da solução $(\mathbf{e}_k, \|\mathbf{w}_k\|)$ do pareto, o SMC-MOBJ foi executado para encontrar soluções com este mesmo erro mas normas diferentes. A figura 3.4 apresenta um grid de soluções para este problema.

É interessante registrar que partir das soluções do pareto para se gerar o grid de soluções, além de ter mais sentido do ponto de vista prático, já que estes pontos tem um significado próprio, tem uma finalidade computacional associada ao algoritmo de modos deslizantes utilizado. Este algoritmo, dependendo do valor de norma e erro que se quer atingir, apresenta problemas de convergência. Assim, partir de soluções do pareto, tende a evitar estes problema, inclusive, reduzindo em muito, o tempo de execução.

3.1 Margem funcional e Margem geométrica

A margem é definida como a distância entre o hiperplano de separação e os pontos de dados mais próximos a ele [34]. Os pontos de dados com os quais se calcula a margem nas máquinas de vetores de suporte (SVM) são chamados de vetores de suporte, e, por serem os mais próximos à superfície

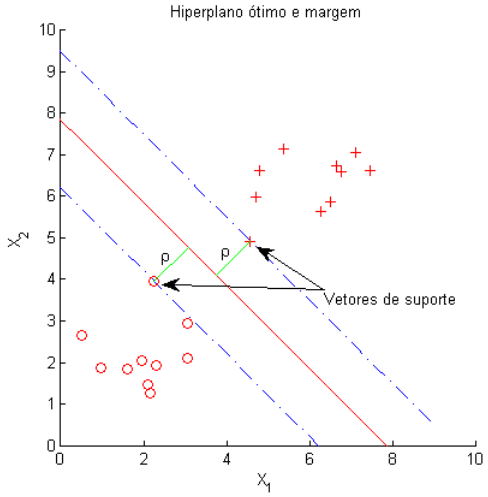


Figura 3.5: Ilustração da idéia de um hiperplano ótimo para padrões linearmente separáveis

de separação eles possuem os maiores erros de classificação. O hiperplano ótimo de separação é aquele que dista equitativamente de ambas as classes. A figura 3.5 exemplifica o conceito de hiperplano ótimo e da margem que é definido como ρ na figura.

A forma de se calcular a margem ρ é dada pela fórmula 3.1 abaixo [55].

$$\rho = \sum_{i=1}^{m} y_i d_i \tag{3.1}$$

onde m é o número de padrões selecionados para se calcular a margem e d é a distância do padrão i ao hiperplano de separação e é calculada conforme a equação 3.2.

$$d_i\left(\mathbf{w}, b, x\right) = \frac{\left(\mathbf{w} \cdot x_i + b\right)}{\|\mathbf{w}\|} \tag{3.2}$$

Geralmente a margem é calculada para os vetores de suporte, pois, em última instância, eles são os pontos mais próximos, de cada classe, ao hiperplano de separação e que sintetizam a informação de todos os padrões da classe para a definição do hiperplano de separação. Em outras palavras, treinar a rede com todos os padrões de entrada ou apenas com os vetores de suporte é equivalente.

Desta forma, o hiperplano ótimo será encontrado ao se maximizar a margem $\rho\left(\mathbf{w}, b\right)$ sujeito a $y_i\left[\left(\mathbf{w} \cdot x_i\right) + b\right] \geq 1$, [32]. O que leva à equação 3.3.

$$\begin{aligned} \rho\left(\mathbf{w}, b\right) &= min_{x_i:y_i=1} d\left(\mathbf{w}, b; X_i\right) + min_{x_j:y_j=-1} d\left(\mathbf{w}, b; X_j\right) \\ \rho\left(\mathbf{w}, b\right) &= min_{x_i:y_i=1} \frac{\left|\mathbf{w} \cdot x_i + b\right|}{\|\mathbf{w}\|} + min_{x_j:y_j=-1} \frac{\left|\mathbf{w} \cdot x_j + b\right|}{\|\mathbf{w}\|} \end{aligned}$$

$$\rho(\mathbf{w}, b) = \frac{2}{\|\mathbf{w}\|} \tag{3.3}$$

Conceitualmente maximizar a margem de separação significa minimizar a norma do vetor de pesos ($\mathbf{w}$) da rede [34] . Contudo, pode-se calcular geometricamente a distância entre o hiperplano de separação e um ponto de dado qualquer. Assim, se for possível determinar, através de algum método, quais padrões de treinamento são os mais relevantes para se determinar o hiperplano de separação ótimo ou apenas a melhor dentre um conjunto de soluções, poder-se-há calcular a distância entre estes pontos e o hiperplano e então definir uma função de custo de maneira que, ao ser minimizada, possa-se encontrar os valores ótimos dos pesos que maximizam a margem geométrica e conseqüentemente determinam o hiperplano ótimo.

Porém, encontrar estes pontos para calcular a margem geométrica não é uma tarefa simples. Não é simples pois não se pode considerar apenas os vetores de suporte como sendo os padrões relevantes ao lidar com um problema semi-supervisionado quando se quer calcular a margem geométrica. Existem alguns detalhes que devem ser considerados ao se calcular e comparar o valor de margem geométrica de um conjunto de soluções.

Retornando ao problema de exemplo,o problema das duas luas , tem-se um grid de soluções geradas com o conjunto de treinamento (indutivo), e se quer definir qual destas soluções generaliza bem o conjunto de validação, levando-se em conta a informação de distribuição dos padrões extraída do conjunto transdutivo. Ou seja, quer-se definir e calcular de alguma maneira a margem geométrica de cada solução do grid, para cada um dos conjuntos de dados (indutivo ρ_l e transdutivo ρ_u), de maneira que sejam comparáveis e que a solução cuja a margem geométrica total (equação 3.4 seja máxima corresponda ao hiperplano ótimo de separação.

$$\rho_{tot} = \rho_l + \rho_u \tag{3.4}$$

A preocupação com a comparabilidade entre os valores de margem geométrica a serem calculados reside exatamente na pré-suposição de se utilizar os vetores de suporte para o seu cálculo. Acontece que dependendo de como os pesos da camada escondida forem escolhidos, os padrões de entrada serão mapeados de uma maneira diferente na camada escondida, e os vetores de suporte podem ser diferentes. É fácil compreender esta afirmação, uma vez, que os pesos do vetor $\mathbf{w}$ dos neurônios da camada escondida para cada uma das soluções do grid (3.4) são diferentes. Com isto, o valor da margem geométrica calculada para cada uma destas soluções não será comparável. Uma primeira solução seria calcular a margem geométrica no espaço de entrada, porém isto seria muito mais difícil se não impossível pois não se conheceria

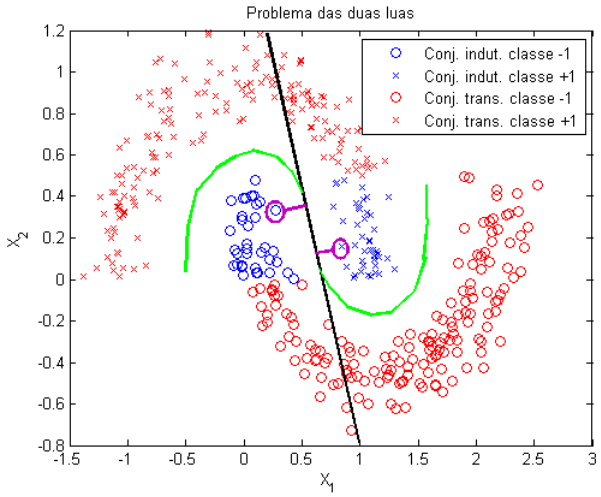


Figura 3.6: A Margem Geométrica para as duas soluções são a mesma se for considerado apenas os vetores de suporte.

os parâmetros que definem o hiperplano no espaço de entrada. Mas mesmo que se implementasse uma maneira, senão direta de se determinar os parâmetros do hiperplano no espaço de entrada, mas mesmo indiretamente, ainda restaria o problema de que para cada solução os vetores de suporte poderiam ser diferentes.

Além disso, duas soluções como as mostradas na figura 3.3 podem ter os mesmos padrões como vetores de suporte, e por conseguinte ter o mesmo valor de margem, porém uma delas é muito melhor que a outra. Tem-se que conseguir caracterizar todo o espaço da separação para poder calcular um valor de margem geométrica que tenha mais sentido.

A idéia então é, de alguma maneira, determinar e fixar os padrões de interesse para resolver o problema. E que padrões seriam estes? Os vetores de suporte não resolvem o problema mas são um bom começo. Poder-se-ia fixar os vetores de suporte de uma das soluções para se calcular a margem geométrica para esta e as demais soluções, mas aí surge mais uma dificuldade. A dificuldade reside no fato de que duas soluções, uma linear que classifica bem o conjunto indutivo e comete muito erros no conjunto transdutivo (reta azul na figura 3.6), e, uma solução não linear que também classifique bem o mesmo conjunto indutivo mas que acerte todas as classificações do conjunto transdutivo (curva vermelha na figura 3.6), podem ter o mesmo valor de margem geométrica se os vetores de suporte escolhidos forem como os circulados de roxo na figura 3.6. Além disso determinar que os vetores de suporte de uma solução serão os mesmo para as demais soluções não faz sentido e se os escolhidos forem de uma solução com alto erro, irá, com certeza, inviabilizar a determinação da separação ótima que atenda aos dois conjuntos de dados.

O que realmente importa quando se leva em consideração a distribuição

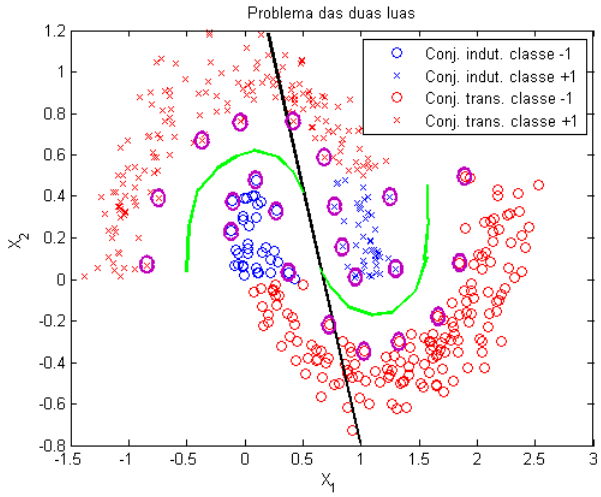


Figura 3.7: Os Padrões circulados em preto são os pontos limites das classes de interesse para o cálculo da margem geométrica.

dos padrões transdutivos para auxiliar a escolha de uma solução gerada pelo conjunto indutivo e que generalize bem na validação é encontrar as soluções que estejam situadas entre as classes, ou seja, que estejam passando pela área de baixa densidade de padrões e, assim, poder determinar qual delas tem a margem geométrica máxima (separação ótima) ou a mais próxima a ela. Sendo assim, os padrões que vão interessar, são aqueles que limitam estas área de baixa densidade, ou seja, aqueles padrões que estão situados ao longo de toda a extensão da região limítrofe entre as classes.

Ao se determinar, de alguma forma, os padrões limites entre as classes, será possível então calcular a margem geométrica para cada solução do grid considerando sempre os mesmo padrões. A maneira como serão selecionados estes limites é objeto do item 3.2.

É importante ressaltar dois pontos importantes: primeiro que alguém pode imaginar que, se, de alguma maneira os pontos limites entre as classes são determinados, já se teria resolvido o problema de classificação, mas isto não é verdade. Em um problema real onde se tem muitos dados sem rótulo, não se pode afirmar, com 100% de certeza, que as regiões de baixa densidade definidas por algum método[1] são realmente os pontos limites. Isto porque em algum momento será necessário definir agrupamentos ou tentar classificar de alguma forma, e assim, pode-se cometer erros. No próprio exemplo usado na descrição dos métodos, poderão ser observados pontos identificados como limites e que não o são pois dentro das classes existem áreas de baixa densidade de padrões e que não estão nas bordas das classes. O outro ponto importante é que se quer calcular uma medida que se preste a indicar se uma solução se encontra no centro da região de separação das classes ou se está

[1] Os métodos para seleção dos pontos limites são objeto de discussão nos ítens 3.2 e 3.3

mais perto de alguma delas, seja em toda a extensão da separação ou em alguma parte..

3.2 Identificando os padrões limites do conjunto transdutivo

3.2.1 O método de identificação de limites por fatiamento

O método de identificação de limites por fatiamento (MILFAT) se baseia no método de agrupamento Fuzzy (Fuzzy c-means clustering) implementado por Jim Bezdek [8]. Esta técnica associa a cada padrão uma probabilidade de pertencer a uma determinada partição. Assim consegue uma maneira de agrupar os padrões em um número predefinido de grupos.

A idéia é bem simples. Aplica-se o método FCM ao conjunto de dados transdutivo para um número final de partições (c) qualquer pré-definido, por exemplo: seis. Ao final do processo, o algoritmo vai retornar, entre outras coisas, uma matriz de probabilidades associando a cada padrão uma probabilidade de pertencer a cada partição.

$$P\left(X_i\right) = \left[p_{i1} \cdots p_{ic}\right] \tag{3.5}$$

Ou seja, ao final do processo, cada padrão X_i vai ter uma probabilidade maior ou menor de pertencer a uma partição. A partição adotada para o padrão em questão será aquela para a qual ele tiver a maior probabilidade. Então, se P_{ic} é a maior probabilidade de X_i ele pertencerá à partição c. Desta forma cada padrão será associado a uma partição. Fazendo $c = 6$, por exemplo, ao final do processamento, os dados do conjunto transdutivo do problema exemplo serão agrupados como na figura 3.8.

Agora que os padrões estão agrupados o próximo passo é identificar os pontos limites de cada uma destas partições. Provavelmente, o agrupamento dos padrões pelo FCM tenderá sempre a guardar alguma coerência com a separação das classes dos dados (que se quer encontrar). Aqui então desenvolveu-se o método do fatiamento no espaço de entrada para tentar identificar o envelope das classes (-1 e +1). A idéia é separar os dados de entrada em faixas e depois verificar onde ocorre a mudança de partição. Com certeza, muitos destes limites indicarão exatamente onde termina e onde começa cada classe. Contudo, também serão encontradas mudanças de partição, mas que não significam mudança de classe. Mais à frente será indicado como contornar este problema.

A figura 3.9 mostra exatamente a idéia que representa o método proposto.

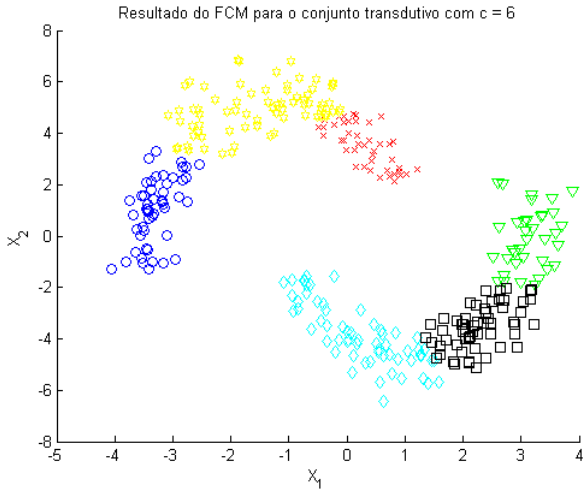


Figura 3.8: Resultado do método FCM para o problema das duas luas transdutivo com 6 clusters

Tem-se um conjunto de dados cujos rótulos (classificação) se supõem desconhecidos e que chamamos de conjunto transdutivo. Este conjunto é composto, no problema exemplo, de 321 padrões de duas variáveis de entrada $X_i = [X_1, X_2]$, portantanto é um problema bidimencional no espaço de entrada. Estes dados são, então, submetidos a um método de agrupamento. Depois, escolhe-se um primeiro eixo, por exemplo, o eixo X_1 e o fatia em f partes. No caso da figura 3.9, f foi definido como sendo oito e a primeira fatia é definida no intervalo $-4 \leq X_1 \leq -3$. Seleciona-se então o segundo eixo (X_2) para ser percorrido do menor valor para o maior, de cada um dos padrões de entrada, verificando onde termina uma partição (gerada pelo FCM) e onde começa a outra, marcando então os padrões onde estas transições ocorrem. Na figura 3.9 está um exemplo dos padrões selecionados para a primeira faixa percorrendo o eixo X_2.

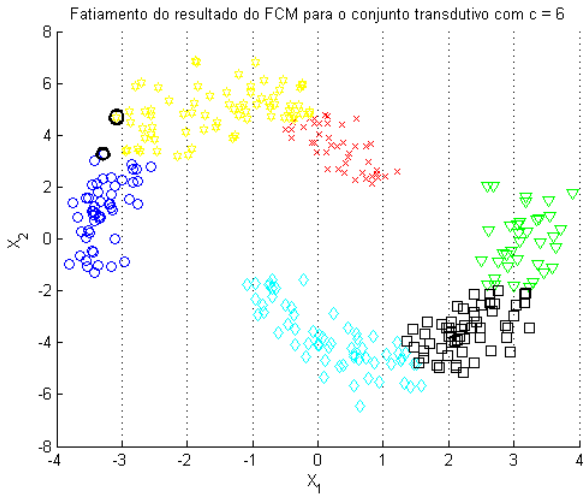


Figura 3.9: Exemplo da aplicação do método de fatiamento no problema das duas luas

Se os padrões de entrada possuissem três dimensões, ou seja, $X_i = [X_1, X_2, X_3]$, seria escolhido primeiro a entrada X_1 para ser fatiada, depois seria percorido os padrões na direção do eixo X_2 marcando as transições das partições do FCM e depois, para a mesma fatia do eixo X_1 seria percorrido os padrões, mas agora na direção do eixo X_3.

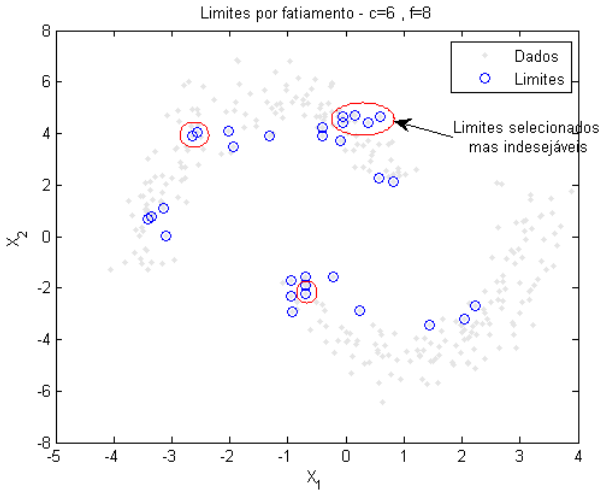


Figura 3.10: Padrões limites entre as partições do FCM para $c = 6$ e $f = 8$

A figura 3.10 mostra como fica o resultado final do fatiamento dos dados do conjunto transdutivo agrupados com o FCM com seis partições. É importante perceber que os padrões envolvidos por uma círculo vermelho são padrões selecionados como limites entre partições do FCM mas que não são de interesse, pois o que se quer caracterizar são os limites da região de mais baixa densidade, ou seja, os limites entre as classes (-1 e +1). A grande questão neste momento é como fazer para eliminar estes padrões indesejáveis, e isto pode ser feito através de uma forma muito simples. Executar-se-há novamente o algoritmo FCM sobre o conjunto de dados transdutivo, só que agora o número de partições c a ser definido deve ser diferente do configurado na primeira vez, por exemplo nove partições. Da mesma forma, ao final do processo, o algoritmo vai retornar uma matriz de probabilidades associando cada padrão a uma das nove classes.

Como feito anteriormente, através da matriz de probabilidades podemos determinar a qual partição o FCM alocou cada padrão de entrada (figura 3.11). Da mesma maneira, aplica-se o fatiamento para cada eixo por vez, caminhando nos demais e marcando as alterações de partições tal qual foi feito na figura 3.10 só que agora com nove partições. O resultado pode ser vista na figura 3.12.

Com estes dois resultados, os padrões limites entre cada uma das seis partições na primeira execução do algoritmo e aqueles padrões selecionados como limites entre cada uma das nove partições na segunda execução, basta veri-

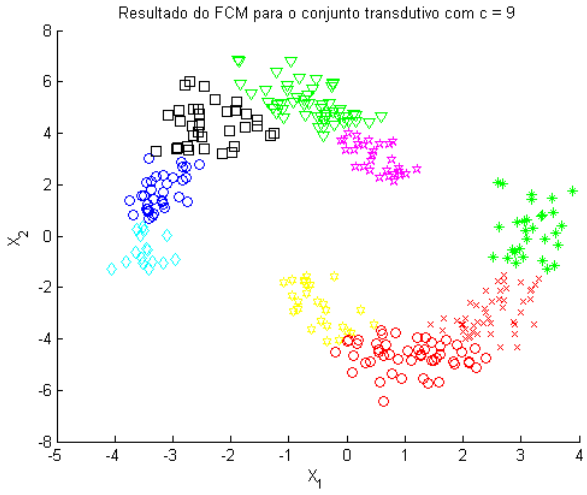


Figura 3.11: Resultado do método FCM para o problema das duas luas transdutivo com 9 clusters

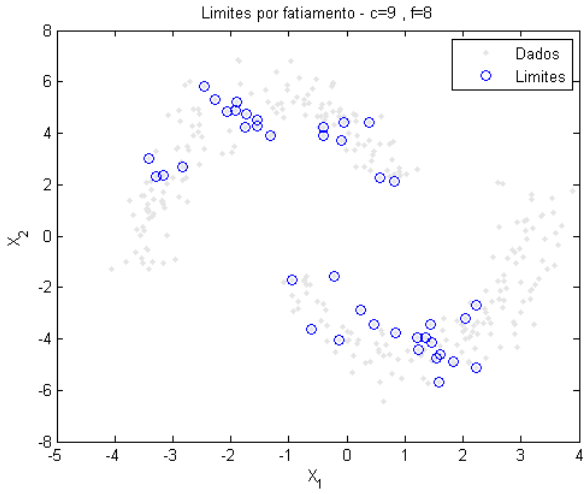


Figura 3.12: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 15$

ficar quais padrões foram selecionados em ambos os casos. Aqueles padrões que estão mais próximos das regiões de baixa densidade tendem a ser selecionados em todas as simulações com partições diferentes, enquanto aqueles que ficam no limite entre as partições mas que são da mesma classe (-1 ou +1) tendem a ser diferentes de uma execução para outra onde o número c de partições é diferente.

A idéia básica é esta, mas na realidade, executa-se este algoritmo para diversos números de partições diferentes, por exemplo, para o problema em questão, poder-se-ia realizar sete simulações onde $6 \leq c \leq 12$.Os limites de número mínimo e máximo de partições vão ser determinados pelo tipo de problema e pelo bom senso. Ao final de todas as simulações, faz-se um histograma, como o da figura 3.13, onde pode-se verificar quais padrões foram selecionados em todas as simulações. A figura 3.14 mostra os padrões que

no histograma da figura 3.13 foram considerados como limites em todas as iterações do FCM.

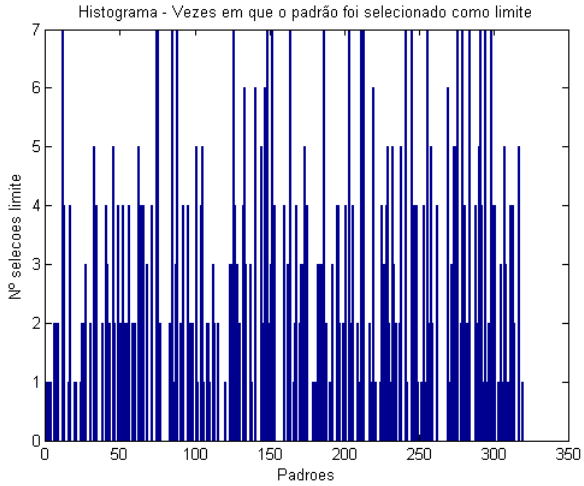


Figura 3.13: Histograma do número de vezes que cada padrão foi selecionado como ponto limite.

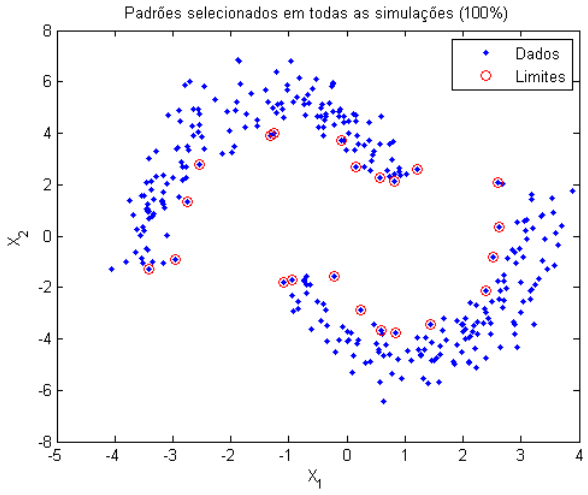


Figura 3.14: Padrões selecionados em todas as simulações (100%)

Também é possível introduzir uma variável de folga k na determinação dos padrões limites. Ao invés de selecionar aqueles que aparecem em todas as simulações, poder-se-ia escolher os padrões que foram selecionados em apenas k% das simulações. Caso se queira considerar como padrões limites apenas aqueles que foram selecionados em todas as sete simulações, faz-se $k = 1$ e os pontos selecionados serão como os da figura 3.14. Já se for considerado $k = 0.8$, serão definidos como limites, os padrões que aparecem selecionados em pelo menos 80% das simulações e os padrões selecionados seriam como os dafigura 3.15. Há que se ter em mente que fazer k menor que 1 pode aumentar o erro no cálculo da margem geométrica, uma vez que padrões que não estão

nas bordas das classes terão maior probabilidade de serem considerados. Na figura 3.15 pode-se observar também este fato.

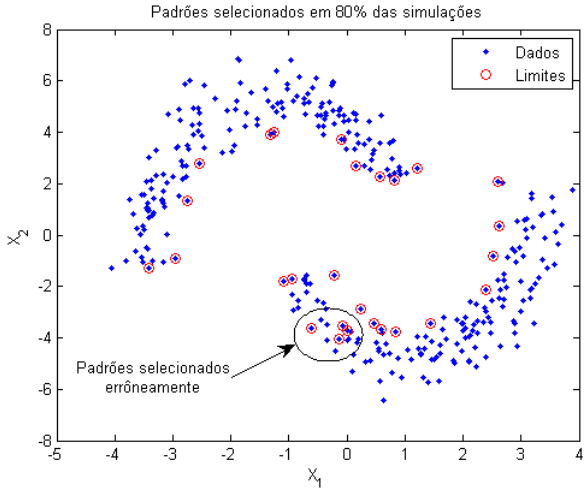


Figura 3.15: Padrões selecionados em 80% das simulações

3.2.2 O método de identificação de limites por probabilidade

O método de identificação de limites por probabilidade (MILP) também se baseia no método de agrupamento Fuzzy (Fuzzy c-means clustering) que, como explicado no ítem 3.2.1, associa a cada padrão uma probabilidade de pertencer a uma determinada partição conseguindo assim uma maneira de agrupar os padrões em um número predefinido de grupos (clusters).

Da mesma maneira que no ítem 3.2.1, a idéia é aplicar o método FCM ao conjunto de dados transdutivo e determinar um número final de partições c qualquer, por exemplo: seis. Ao final do processo, o algoritmo vai retornar, entre outras coisas, uma matriz de probabilidades associando a cada padrão uma probabilidade de pertencer a cada partição.

$$P\left(X_i\right) = \left[p_{i1} \cdots p_{ic}\right] \tag{3.6}$$

Se a maior probabilidade do padrão X_i pertencer a uma classe for a associada à classe c, ou seja p_{ic} é a maior probabilidade de X_i, então ele pertencerá à classe c (figura 3.16). Contudo, tomando as n menores probabilidades de todos os padrões que ficaram em uma determinada classe c por exemplo, teremos os n elementos mais distantes do centro do cluster definido no método FCM, ou seja, em tese, os elementos limites destas partições. A figura 3.17 mostra um exemplo para $n = 15$.

Com certeza só isto não garante que os n dados selecionados estarão exatamente na divisão entre as classes reais do problema, e além do mais, quanto

maior o n maior poderá ser o número de padrões selecionados mais internamente à partição (ver figuras 3.18 e 3.19). Contudo, se n for bem ajustado, muito provavelmente, muitos deles estarão. Daí surge a mesma questão que no ítem 3.2.1: como eliminar os padrões selecionados como limites mas que estão na verdade dentro da mesma classe real (-1 ou +1)?

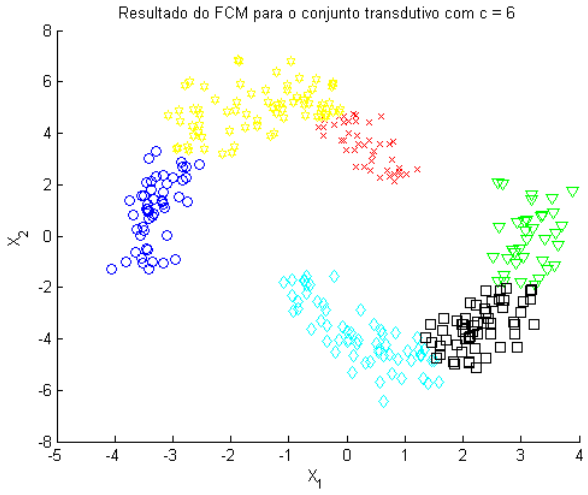


Figura 3.16: Resultado do método FCM para o problema das duas luas transdutivo com 6 clusters

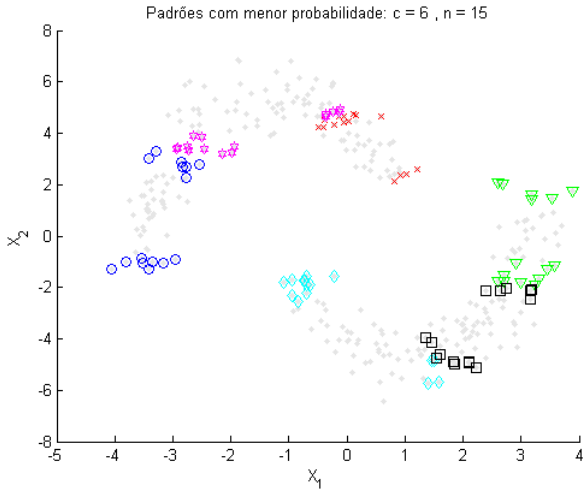


Figura 3.17: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 15$

Da mesma forma que no método MILFAT executar-se-há novamente o algoritmo FCM sobre o conjunto de dados transdutivo, só que agora o número de partições c a ser definido deve ser diferente do configurado na primeira vez, por exemplo nove partições. Ao final do processo, o algoritmo vai retornar uma matriz de probabilidades associando cada padrão a uma das nove classes.

Como feito anteriormente, através da matriz de probabilidades podemos determinar a qual partição o FCM alocou cada padrão de entrada (figura 3.20).

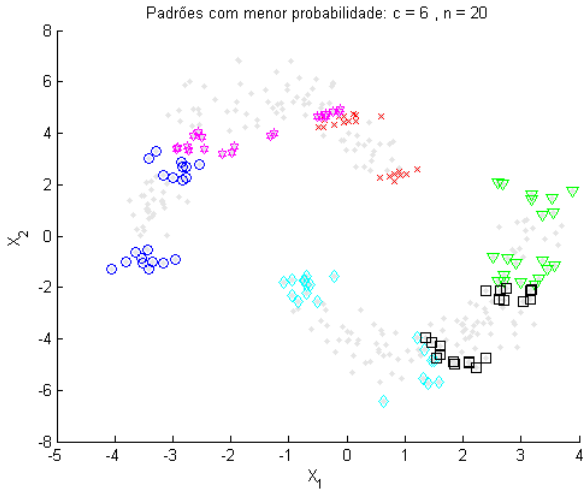


Figura 3.18: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 20$

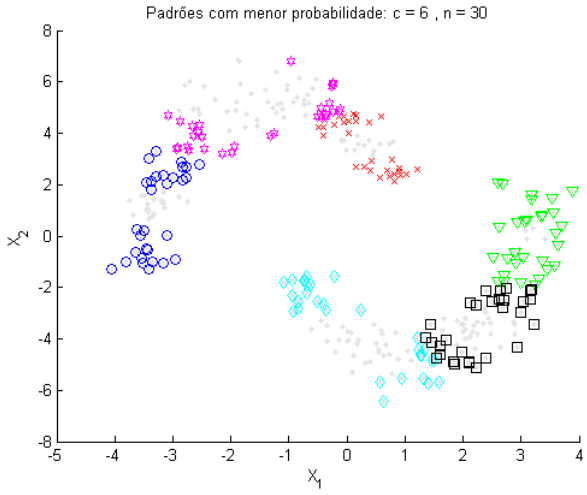


Figura 3.19: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 30$

Da mesma maneira, seleciona-se os n padrões que apresentam a menor probabilidade dentro de uma mesma partição c. Os resultados desta última operação podem ser visto nas figuras 3.21, 3.22 e 3.23, para diferentes valores de n.

Como no método MILFAT, tendo estes dois resultados, os n padrões de menor probabilidade de pertencer a cada uma das seis partições na primeira execução do algoritmo e os n padrões de menor probabilidade de pertencer a cada uma das nove partições na segunda execução, basta verificar quais padrões foram selecionados em ambos. Aqueles padrões que estão mais próximos das regiões de baixa densidade tendem a ser selecionados em todas as simulações com partições diferentes, enquanto aqueles que ficam no limite entre as partições mas que são da mesma classe (-1 ou +1) tendem a ser diferentes.

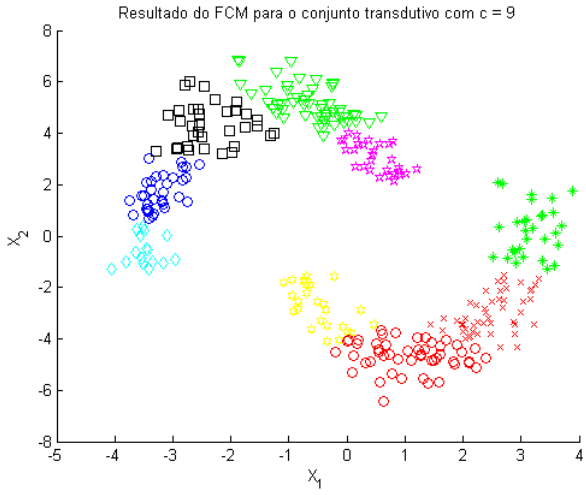


Figura 3.20: Resultado do método FCM para o problema das duas luas transdutivo com 9 clusters

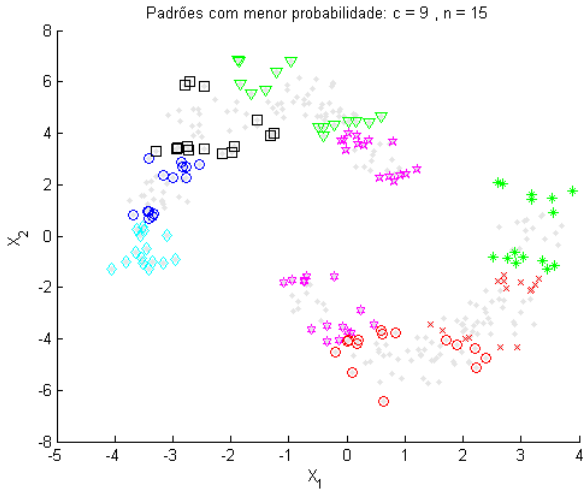


Figura 3.21: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 15$

Pode-se executar este algoritmo para diversos números de partições diferentes, por exemplo, variando de duas partições até dez. Os limites de número mínimo e máximo de partições vão ser determinados pelo tipo de problema e pelo bom senso. Ao final de todas as simulações, faz-se um histograma, como o da figura 3.24, onde pode-se verificar quais padrões foram selecionados em todas as simulações (no caso, apenas duas simulações). A figura 3.25 mostra os padrões que foram selecionados como limites para todas as simulações do FCM.

Como no método MILFAT também é possível introduzir uma variável de folga k na determinação dos padrões limites. Ao invés de selecionar aqueles que aparecem em todas as simulações, pode-se escolher os padrões que foram selecionados em k% das simulações. Para o exemplo em questão, supondo que fossem realizadas sete simulações onde $6 \leq c \leq 12$, o histograma resultante

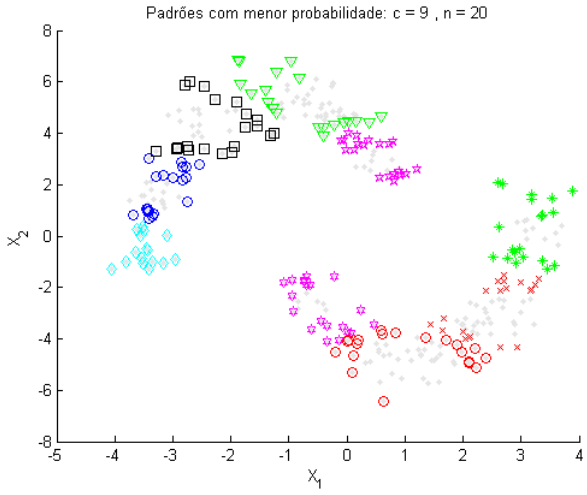


Figura 3.22: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 20$

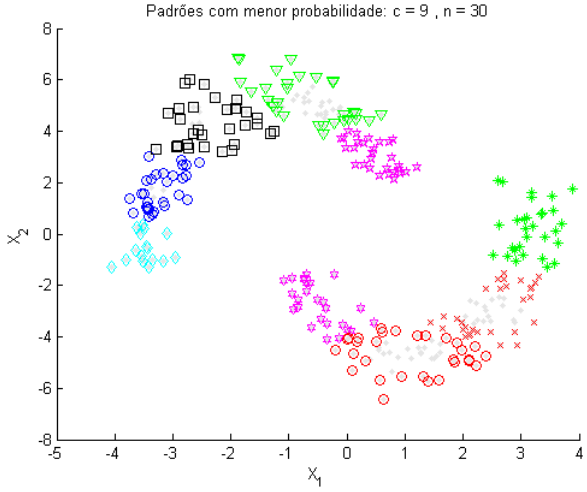


Figura 3.23: Seleção dos n elementos de mais baixa probabilidade de cada partição do FCM com $n = 30$

seria o da figura 3.24. Caso se queira considerar como padrões limites apenas aqueles que foram selecionados em todas as sete simulações, faz-se $k = 1$ e os pontos selecionados serão os da figura 3.25. Já se for considerado $k = 0.8$, serão definidos como limites, os padrões que aparecem selecionados em 80% das simulações (figura 3.26). Há que se ter em mente que fazer k menor que 1 pode aumentar o erro no cálculo da margem geométrica, uma vez que padrões que não estão nas bordas das classes terão maior probabilidade de serem considerados.

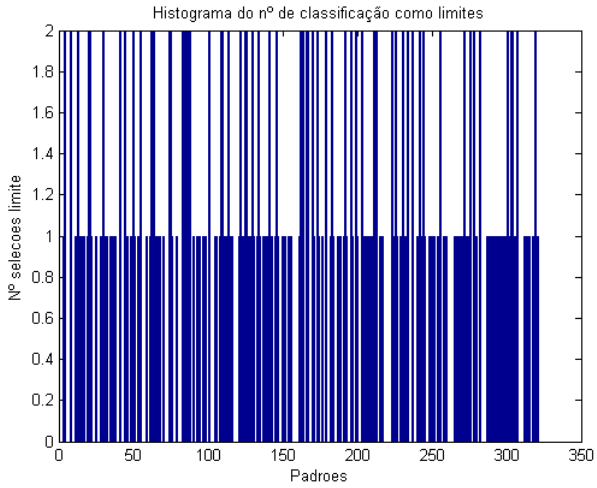


Figura 3.24: Histograma do número de vezes que cada padrão foi selecionado como ponto limite.

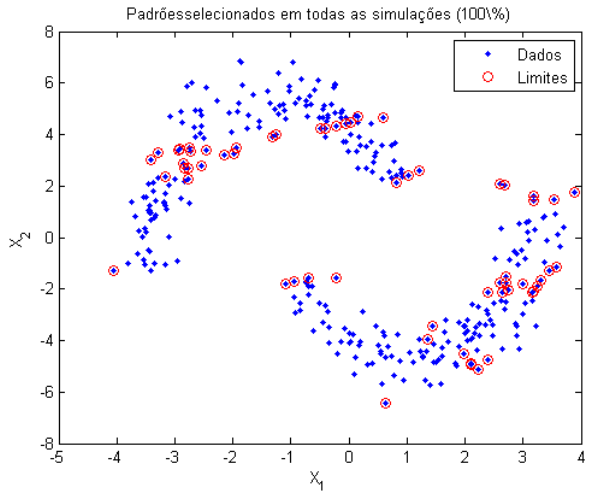


Figura 3.25: Padrões selecionados em todas as simulações (100%)

3.3 Identificando os padrões limites do conjunto Indutivo

Para a seleção dos padrões a serem considerados como sendo limites entre as classes, ou seja, limites das regiões de baixa densidade em termos quantitativos para o conjunto indutivo, faz-se o processo de fatiamento descrito em 3.2.1, porém aqui não há a necessidade de se utilizar o artifício de gerar clusters com o FCM para tentar aproximar o limite das classes. Como se conhece previamente a classificação deste conjunto de dados basta identificar onde há a mudança de uma classe para outra.

Este método, o MILCL (Método de Identificaçao dos Limites de Classes *Labeled*), consiste em fatiar por vez cada variável de entrada, e verificar nas demais onde ocorre a mudança de classe ao se "caminhar" ao longo dos valores

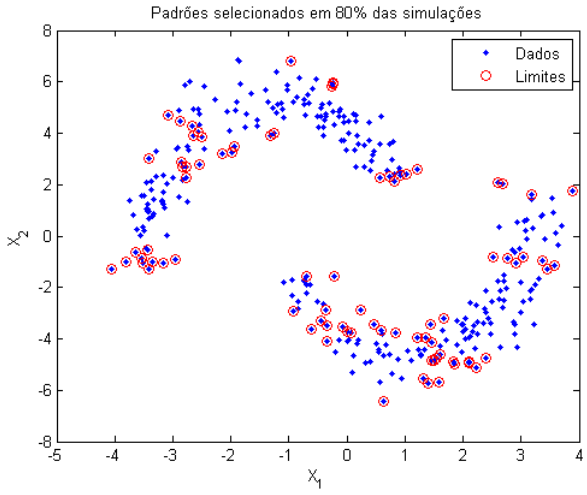


Figura 3.26: Padrões selecionados em 80% das simulações

ordenados.

Cada variável de entrada pode ser traduzida como um eixo no espaço de entrada. Para o problema das duas luas, como exemplificado na figura 3.9 da seção 3.2.1, seleciona-se o primeiro eixo X_1, ordena-se seus valores e separa-se em f faixas. Toma-se os padrões que tem valores de X_1 dentro da primeira faixa, por exemplo, e ordena-se os valores do eixo X_2. Depois disto "caminha-se" pelos padrões na direção do menor valor de X_2 para o maior valor, verificando se houve mudança de classe. Quando ocorre a mudança de classe registra-se quais os padrões onde ocorreu. A figura 3.27 mostra o resultado da aplicação do método para o problema exemplo das duas luas com $f = 5$.

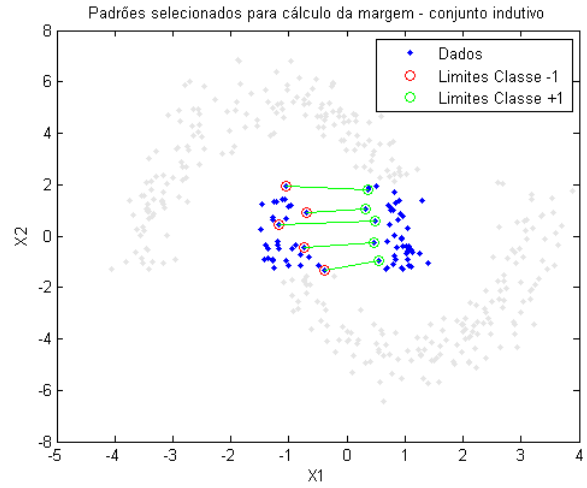


Figura 3.27: Padrões selecionados como limites do conjunto indutivo

3.4 O cálculo da Margem geométrica

Tem-se então neste momento os padrões que devem ser considerados para o cálculo da margem geométrica. Quer-se determinar uma forma de calcular a distancia entre o hiperplano de separação na camada escondida e os padrões de entrada ao longo de todo o limite de separação entre as classes. As razões para tanto foram discutidas no ítem 3.1 desta dissertação.

A equação 3.1 calcula a margem como sendo o somatório da distância ponderado pelo pela saída da rede y. Neste caso, para o conjunto indutivo, por exemplo, para soluções que classificam corretamente todos os pontos, o valor da margem é maior que o valor para as soluções que não classificam corretamente todos os pontos. Porém, quando se compara o valor da margem para soluções que classificam corretamente todos os padrões de entrada do conjunto indutivo, não se pode definir qual a melhor solução. Não apenas baseado neste valor calculado desta maneira, e a razão para tanto pode ser o balanceamento das classes dos dados.

Só para simplificar a análise, seja um conjunto indutivo composto por apenas três padrões. Um da classe (-1), portanto seu y de saída é (-1) e os outros dois da classe (+1) com $y = +1$. Sejam também três soluções que classificam corretamente o conjunto (veja figuras 3.28 e 3.29).

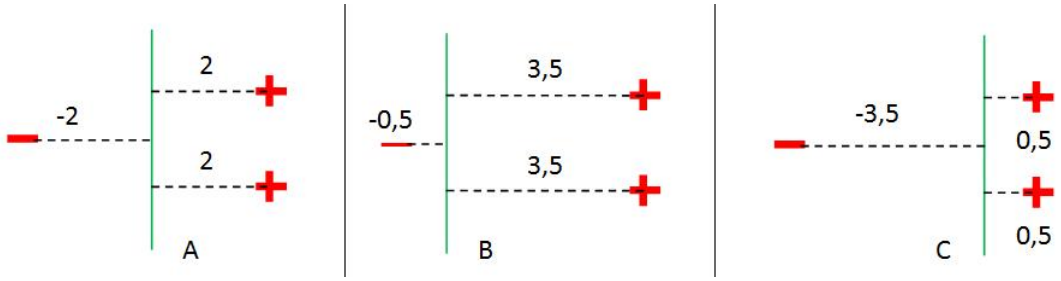


Figura 3.28: Problema do desbalanceamento das classes no cálculo da margem como soma

A margem como calculada na equação 3.1, para cada solução na figura 3.28 fica da seguinte forma:

$$\begin{aligned} \rho_A &= (-2)(-1) + (2)(1) + (2)(1) = 6 \\ \rho_B &= 7,5 \\ \rho_C &= 7,5 \end{aligned}$$

Onde ρ_A, ρ_B e ρ_C são as margens calculadas para cada um dos exemplos

da figura 3.29.

A melhor solução é a mostrada na figura 3.29A pois ela maximiza a margem, contudo seu valor de margem calculado desta maneira não é nem o mínimo nem o máximo. Isso, considerando que o mapeamento da camada escondida seja o mesmo nos três casos acima. Caso o mapeamento desta camada for diferente, como discutido anteriormente estes resultados ficam incomparáveis.

Da mesma forma, se o número de padrões considerados no cálculo da margem geométrica em uma classe e na outra forem exatamente o mesmo, somar as distâncias ponderado pela saída da rede também pode não ajudar muito.

Seja a situação proposta na figura 3.29.

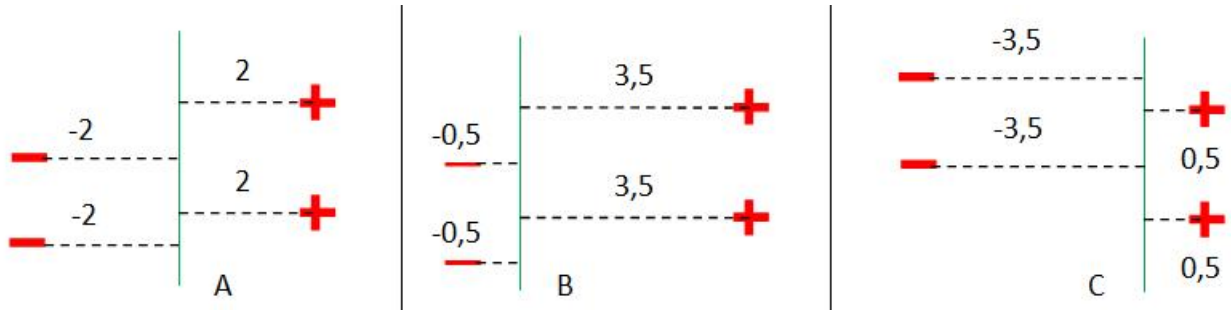


Figura 3.29: Problema do balanceamento das classes no cálculo da margem como soma

Se por hipótese, os padrões das duas classes fossem simétricos e balanceados, a margem geométrica teria rigorosamente o mesmo valor para qualquer solução que classifique corretamente os padrões:

$$\begin{aligned} \rho_A &= (-2)(-1) + (-2)(-1) + (2)(1) + (2)(1) = 8 \\ \rho_B &= 8 \\ \rho_C &= 8 \end{aligned}$$

De qualquer forma, o número de pontos em uma classe e na outra, assim como sua posição podem alterar significativamente o resultado da margem da forma como ela é calculada na equação 3.1.

Mas e se ao invés de somar as distâncias elas fossem multiplicadas?

Para o problema proposto na figura 3.28 tem-se:

$$\rho_A = [(-2)(-1)]X[(2)(1) + (2)(1)] = 8$$

$$\begin{aligned}\rho_B &= 3,5\\ \rho_C &= 3,5\end{aligned}$$

Para o problema proposto na figura 3.29 tem-se:

$$\begin{aligned}\rho_A &= [(-2)(-1)+(-2)(-1)]*[(2)(1)+(2)(1)] = 16\\ \rho_B &= [(-0,5)(-1)+(-0,5)(-1)]*[(3,5)(1)+(3,5)(1)] = 7\\ \rho_C &= 7\end{aligned}$$

Assim, os resultados indicariam que a solução A, de margem máxima, é a solução ótima. A função de cálculo da margem geométrica seria então:

$$\rho = \rho_1 * \rho_2 \tag{3.7}$$

onde ρ_1 é a soma das distâncias dos padrões da classe -1 ao hiperplano ponderadas pela saída da rede e ρ_2 é a soma das distâncias dos padrões da classe +1 ao hiperplano também ponderadas pela saída da rede.

Mas ainda resta o problema dos diferentes mapeamentos para a camada escondida. Cada solução do grid da figura 3.4 apresentado no início deste capítulo, apresenta pesos da camada escondidas diferentes uns dos outros, o que resulta em diferentes mapeamentos. E a pergunta que se faz é a seguinte? Será que isto altera o valor relativo de margem geométrica de cada uma destas soluções. Com certeza o valor absoluto pode ser bem diferente, mas será que a solução que tem maior margem vai continuar a ter o maior valor para qualquer mapeamento? Provavelmente não, e sendo assim, buscou-se uma terceira forma de se calcular a margem geométrica de maneira a evitar a influência do mapeamento da camada escondida.

Logo se imaginou que a solução então seria calcular a margem geometricamente no espaço de entrada pois assim evitar-se-ia a influência do mapeamento dos dados de entrada na camada escondida. A idéia seria mapear no espaço de entrada a superfície de separação definida pela solução do grid na camada escondida. Com certeza, não seria fácil, em princípio, descobrir os parâmetros que definem a superfíce de separação no espaço de entrada, contudo, poder-se-ia definir facilmente, diversos pontos nesta superfície de tal modo que fosse possível calcular a distância dos pontos limites das classes aos pontos desta superfície de separação. Mas uma segunda idéia parece ser mais simples e ao mesmo tempo suficiente para se calcular a margem geométrica

na camada escondida, eliminando o efeito dos diferentes mapeamentos.

Na verdade é importante ter em mente que quanto mais ao centro da região entre as classes estiver um hiperplano de separação, mais próximo será o valor da distância deste hiperplano à classe -1, do valor da distância do hiperplano à classe +1. Em última instância, pode-se considerar ρ_1 como sendo a distância do hiperplano à classe -1 e ρ_2 como a distância à classe +1, então, se $\rho_1/\rho_2 = 1$ o hiperplano está exatamente no centro entre as duas classes. Para garantir que ρ máximo seja 1 garante-se que no denominador da equação 3.8 esteja sempre a maior distância.

$$\rho = \frac{min\,(\rho_1, \rho_2)}{max\,(\rho_1, \rho_2)} \tag{3.8}$$

3.5 Seleção da melhor solução

Agora só falta determinar, dentre cada uma das soluções do conjunto indutivo relacionadas no grid gerado, qual delas atende bem tanto o conjunto de treinamento como o transdutivo (validação).Para cada uma das soluções, soma-se a margem geométrica calculada para o conjunto indutivo com a calculada para o transdutivo e depois escolhe-se aquela de maior margem total, ou seja, escolher-se-á aquela solução $\mathbf{G} : (e^*, \mathbf{w}^*)$ do grid que, maximiza a margem geométrica total ρ_{tot} que é dada pela fórmula 3.9.

$$\rho tot = \rho_{indutivo} + \rho_{transdutivo} \tag{3.9}$$

O cálculo de $\rho_{indutivo}$ e $\rho_{transdutivo}$ é detalhado na seção 3.4.

3.6 Considerações finais

A principal discussão deste capítulo é a definição e extração de uma informação do conjunto de dados transdutivo que possibilite ou ao menos ajude a definir soluções que classifiquem bem os padrões de treinamento e os de validação. A informção extraida é a margem geométrica. Aqui foi discutido como calculá-la de forma mais apropriada. Neste cálculo é necessário definir e configurar alguns parâmetros, sobretudo a quantidade de partições a serem consideradas no FCM e a quantidade de pontos limites a serem selecionados, que serão discutidos mais amplamente na conclusão deste trabalho, contudo os resultados são promissores, como poderá ser verificado no capítulo 5.

CAPÍTULO 4

Os dados do problema de aplicação

Os dados que serão utilizados para aplicar o método criado constituem-se pelas informações de diversos genes de 133 pacientes que foram diagnosticados com câncer de mama. Estas informações são obtidas através da tecnologia de *microarrays* que permite analisar ao mesmo tempo uma enorme quantidade de genes. Baseado nisso e querendo garantir uma boa precisão destas informações de cada paciente, utilizou-se mais de uma sonda para caracterizar cada gene. Tudo isso resulta em uma grande massa de dados que precisa ser analisada. Ao todo, foram geradas as expressões de mais de 22.000 sondas para cada um dos pacientes estudados [36].

Estes pacientes foram submetidos ao tratamento quimeoterápico e ao final, verificou-se quantos deles ainda apresentavam células cancerígenas na mama. Os pacientes cujo tratamento não conseguiu eliminar todas as células cancerígenas foram classificados como RD (residual disease), que neste trabalho serão classificados como NOPCR, e os que conseguiram eliminar foram classificados como pCR (pathologic complete response) que aqui também serão classificados como PCR. O objetivo então é utilizar a informação genética destes pacientes para tentar identificar se novos pacientes, se submetidos ao tratamento, seriam prováveis NOPCR ou PCR.

Os dados dos 133 pacientes foram coletados nos Estados Unidos, França e uma pequena parte no Peru. Destes, 82 foram coletados em Huston, EUA, e serão utilizados como o conjunto de dados indutivo ou de treinamento. Já os dados de 51 pacientes foram coletados na França em Villejuif e no Peru. Estes comporão o conjunto de dados transdutivos e de validação.

Acontece que uma base de dados com mais de 22.000 informações para cada um dos 133 pacientes é um volume de dados muito grande e de difícil

análise, sem contar que muitas das expressões das sonda obtidas são redundantes já que se queria garantir a representatividade dos genes. Diversas técnicas então foram aplicadas na intenção de reduzir este conjunto de dados e tentar selecionar aquelas sondas mais relevantes ao problema. Nesta dissertação vai-se trabalhar com dois conjuntos de sondas selecionadas por Horta em [36]. Um conjunto contém 18 sondas e o outro contém 11 sondas. Não necessariamente as sondas dos dois conjuntos são diferentes, pois há intercessão entre estes dois conjuntos. Em [36] pode-se encontrar informações mais detalhadas sobre os métodos de seleção das sondas, contudo, neste capítulo, serão explicitadas as técnicas abordadas no trabalho citado.

4.1 Seleção das sondas

Horta em sua dissertação de mestrado [36] cita o trabalho de Hess [35] na seleção das sondas. Eles fizeram a seleção baseados em ranking do p-valores resultantes da aplicação do teste-t. Basicamente eles aplicaram o teste-t no conjunto de dados de nível de expressão que foram previamente convertidos para uma escala logaritmica (log_{10}), e depois ordenaram os genes pelo ranking de p-valores. Segundo este critério, os genes com mais informações relevantes foram utilizados para criar previsores multigênicos usando os classificadores SVM [65] [54], KNN e DLDA [57]. Daí, para cada combinação de genes, realizou-se a validação cruzada e avaliou-se a performance de classificação através da área acima da curva ROC (AAC) definindo que o classificador de melhor performance era o DLDA composto por 30 sondas que são mostradas na tabela 4.1 que foi retirado do trabalho de Horta.

O segundo método de seleção de sondas descrito por Horta em seu trabalho é a seleção baseada em intervalos de níveis de expressão, que foi aplicada por Natowicz et al. [45] ao banco de dados. Como Horta bem explicita, este método é bastante simples e até mesmo intuitiva. Basicamente consiste em considerar que cada sonda s é um classificador independente capaz de classificar se o paciente é PCR ou NOPCR. Depois disso, calcula-se a média e o desvio padrão dos valores de expressão de cada sonda para o conjunto de pacientes PCR e para os NOPCR. Ou seja, Para todos os pacientes classificados como PCR, calcula-se a média e o desvio padrão dos valores de expressão de cada sonda e faz-se o mesmo para o grupo NOPCR. Estes intervalos definidos pela média e desvio padrão de cada grupo para cada sonda podem apresentar intercessão ou não. Todo valor de expressão fora dos intervalos ou na intercessão dos mesmos são considerados como indefinidos e desconsiderados no processo.

Ao determinar estes intervalos para cada sonda, calcula-se, utilizando o conjunto de treinamento, a capacidade de classificação de cada sonda $V(s)$ que

Tabela 4.1: Sondas selecionadas por Hess. *Fonte: Horta [36]*

Sonda	Gene		Sonda	Gene
203929_s_at	MAPT		202204_s_at	AMFR
203930_s_at	MAPT		209617_s_at	CTNND2
212745_s_at	BBS4		205354_at	GAMT
203928_x_at	MAPT		204509_at	CA12
212207_at	THRAP2		214124_x_at	FGFR1OP
217542_at	MBTPS1		213234_at	KIAA1467
206401_s_at	MAPT		219051_x_at	METRN
215304_at	PDGFRA		219044_at	FLJ10916
219741_x_at	ZNF552		203693_s_at	E2F3
204916_at	RAMP1		214053_at	ERBB4
208945_s_at	BECN1		215616_s_at	JMJD2B
213134_x_at	BTG3		209773_s_at	RRM2
219197_s_at	SCUBE2		219438_at	FLJ12650
204825_at	MELK		205696_s_at	GFRA1
205548_s_at	BTG3		201508_at	IGFBP4

é função do número de pacientes PCR classificados corretamente, do número real de pacientes da classe PCR, do número de pacientes NOPCR classificados corretamente e do número real de pacientes da classe NOPCR. As expressões e maiores detalhes podem ser encontrados em [36].

Utilizando o voto majoritário das sondas, procurou-se selecionar as sondas que obtivessem uma boa classificação dos conjuntos de treinamento e de validação. A tabela 4.2 mostra as 30 sondas selecionadas no trabalho de Natowicz et al. [45] e que foram listadas no trabalho de Horta.

O terceiro método de seleção de sondas foi o utilizado por Horta em seu trabalho para definir um grupo de 18 sondas. Em [36] Horta descreve detalhadamente como chegou a este conjunto reduzido de sondas e mostra também que seus resultados são compatíveis com os dos demais métodos, e em certos aspectos até melhor. Contudo, nesta dissertação, far-se-há apenas uma explicação geral sobre o método.

O volcano plot é um gráfico dos p-valores calculados pelo teste-t de cada sonda, pelo *fold-change* [20] calculado utilizando o log_2 da média das razões entre as expressões dos genes de cada classe. A técnica do *fold-change* consiste em avaliar o logaritmo da média entre duas condições (ou a média entre as razões de todos as amostras) e considerar todos os genes que diferem mais do que um valor de corte arbitrário verificando-se assim, se a expressão do gene sobre uma condição ou sobre uma classe é um certo número de vezes maior ou menor que o valor da expressão do mesmo sobre outra condição ou em outra classe [36].

Desta forma, utilizando o volcano plot, conseguiu-se selecionar um con-

Tabela 4.2: Sondas selecionadas por Natowicz. *Fonte: Horta [36]*

Sonda	Gene		Sonda	Gene
213134_x_at	BTG3		204825_at	MELK
205548_s_at	BTG3		215867_x_at	CA12
209604_s_at	GATA3		214164_x_at	CA12
209603_at	GATA3		212046_x_at	MAPK3
212207_at	THRAP2		209602_s_at	GATA3
201826_s_at	SCCPDH		212745_s_at	BBS4
205339_at	SIL		203139_at	DAPK1
209016_s_at	KRT7		203226_s_at	SAS
201755_at	MCM5		219044_at	FLJ10916
204862_s_at	NME3		203693_s_at	E2F3
219051_x_at	METRN		220016_at	AHNAK
211302_s_at	PDE4B		214383_x_at	KLHDC3
212660_at	PHF15		212721_at	SFRS12
200891_s_at	SSR1		202200_s_at	SRPK1
202392_s_at	PISD		217028_at	CXCR4

junto de sonda mais relevantes e utilizando os classificadores Naïve Bayes e o classificar de voto majoritário devido à simplicidade de ambos, e analisando a área acima da curva ROC calculado com os resultados dos classificadores, pôde-se chegar a um conjunto de 18 sondas mais representativas para o problema. Maiores detalhes sobre a seleção destas 18 sondas podem ser encontradas em [36]. A tabela 4.3 mostra as sondas selecionadas por este método.

Tabela 4.3: 18 Sondas selecionadas por Horta. *Fonte: Horta [36]*

Sonda	Gene		Sonda	Gene
205548_s_at	BTG3		214164_x_at	CA12
204825_at	MELK		205044_at	GABRP
204913_s_at	SOX11		208370_s_at	DSCR1
219051_x_at	METRN		220559_at	EN1
210147_at	ART3		202342_s_at	TRIM2
217028_at	CXCR4		217838_s_at	EVL
201508_at	IGFBP4		203628_at	IGF1R
203929_s_at	MAPT		203789_s_at	SEMA3C
205225_at	ESR1		221728_x_at	XIST

Horta ainda descreve em seu trabalho [36] resultados obtidos com a utilização de 18 sondas e o modelo de LASSO (*pruning*) que permitiu reduzir o espaço de entrada para 11 sondas e que também serão utilizadas na aplicação do modelo semi-supervisionado ao problema proposto. A tabela 4.4 mostra as 11 sondas que restaram.

Tabela 4.4: 11 Sondas selecionadas por Horta. *Fonte: Horta [36]*

Sonda	Gene		Sonda	Gene
205548_s_at	BTG3		214164_x_at	CA12
204913_s_at	SOX11		202342_s_at	TRIM2
219051_x_at	METRN		220559_at	EN1
217028_at	CXCR4		203929_s_at	MAPT
205225_at	ESR1		221728_x_at	XIST
204825_at	MELK			

4.2 Considerações finais

Para a aplicação do método semi-supervisionado MSMG descrito no capítulo 3 serão utilizados o conjunto de dados formados pelas 30 sondas selecionadas por Natowicz e o conjunto de dados composto pelas 18 sondas selecionados por Horta, bem como seu subconjunto de 11 sondas. Não é intenção aqui atestar a qualidade ou a capacidade de representação destes conjuntos de dados, mesmo porque isto foi muito bem feito em [36]. Contudo, como o método é semi-supervisionado, o que se quer é verificar se as características de margem geométrica extraídas do conjunto transdutivo destes grupos de sondas cada vez mais reduzidos é capaz de selecionar bem soluções que atendam tanto ao conjunto indutivo quanto ao transdutivo. Ou seja, verificar como este método se comportará com a escassez ainda maior de dados, uma vez que o número de pacientes já é reduzido.

CAPÍTULO 5

Resultados - Aplicação dos métodos ao problema

Apresenta-se neste capítulo os resultados obtidos na aplicação do método semi-supervisionado baseado na margem geométrica ao conjunto de dados selecionados. São descritas as etapas com seus resultados parciais e ao final, os resultados obtidos são comparados com os apresentados no trabalho de Horta [36].

Como discutido no capítulo 4, os dados coletados são informações de níveis de expressão de genes com mais de 22000 sondas para cada um dos pacientes estudados. Fez-se necessário então, através de alguns métodos, selecionar aquelas sondas que melhor caracterizassem a classificação dos pacientes nas duas classes (PCR, NOPCR). A seção 4.1 descreve como foram feitas as seleções destas sondas mais representativas. Para a avaliação do método semi-supervisionado descrito no capítulo 3 escolheu-se alguns destes conjuntos de sondas:

- o conjunto de 30 sondas selecionadas por Natowicz;
- o conjunto de 18 sondas selecionadas por Horta;
- o conjunto de 11 sondas selecionadas por Horta.

5.1 O grid de soluções

Primeiramente, como detalhado no capítulo 3, utilizando os 82 padroes do conjunto indutivo como sendo os dados de entrada de um algoritmo multi-objetivo gerou-se a curva do pareto para cada um dos conjuntos de sondas

definidos. Em todos estes conjuntos, a topologia de rede utilizada foi uma MLP com uma cada escondida composta de dez neurônios e com um neurônio na camada de saída. A figura 5.1 mostra os paretos gerados para cada um dos conjuntos de sondas.

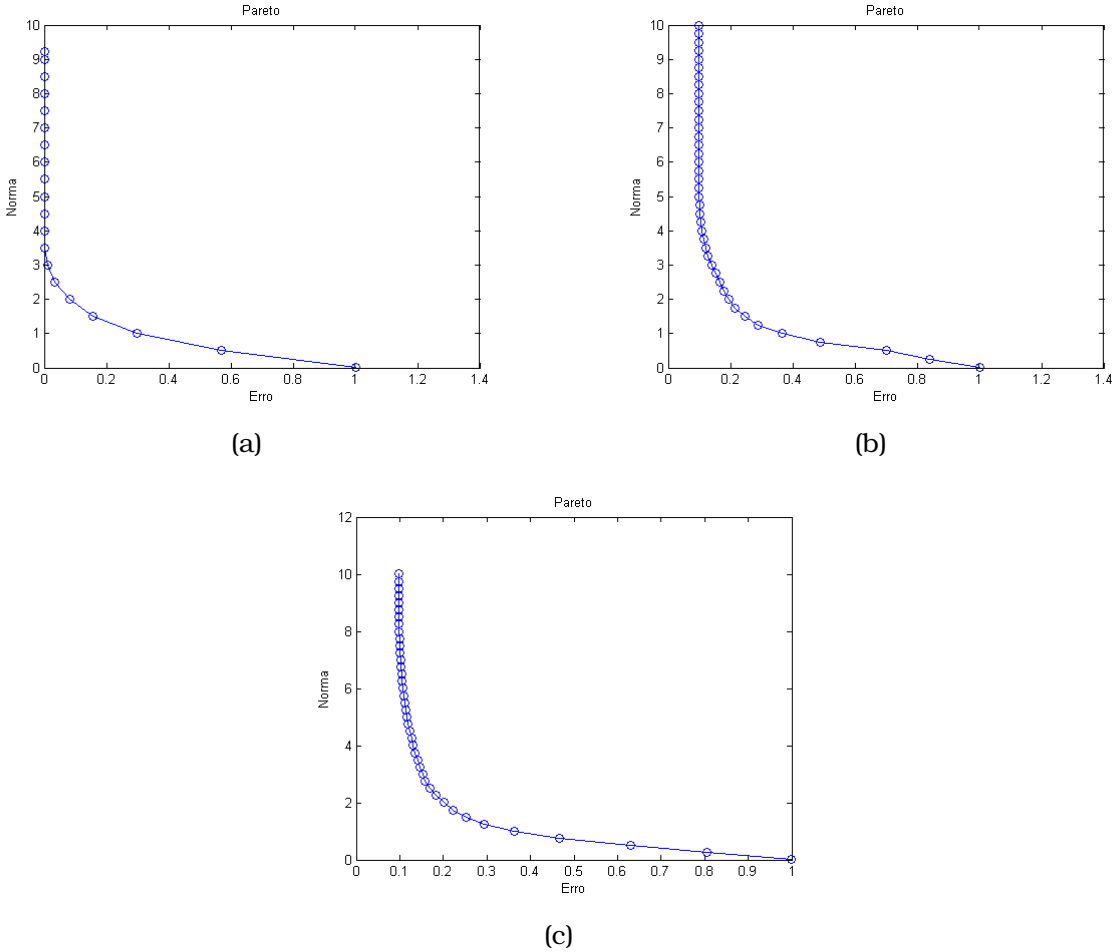


Figura 5.1: Paretos gerados para cada um dos conjuntos de sondas selecionados. (a) conjunto de 30 sondas de Natowicz, (b) conjunto de 18 sondas e (c) conjuntos de 11 sondas.

A partir dos paretos de cada um dos conjuntos de sondas selecionados, gerou-se então um grid de soluções utilizando o algoritmo de modos deslizantes [16]. Vale lembrar que para gerar o grid de soluções, ainda está-se utilizando o conjunto de dados indutivo apenas. O pareto será utilizado como uma melhor condição inicial ao algoritmo de modos deslizantes para tentar evitar problemas de convergência e assim conseguir chegar aos pontos desejados de norma e erro e montar o grid mais rapidamente inclusive. A figura 5.2 mostra os grids encontrados para cada conjunto de sondas.

O grid mostrado na figura 5.2(a) relativo ao conjunto indutivo de 30 sondas foi gerado com um menor número de soluções devido ao grande número de variáveis de entrada. Isto para que o tempo de processamento não ficasse muito demorado.

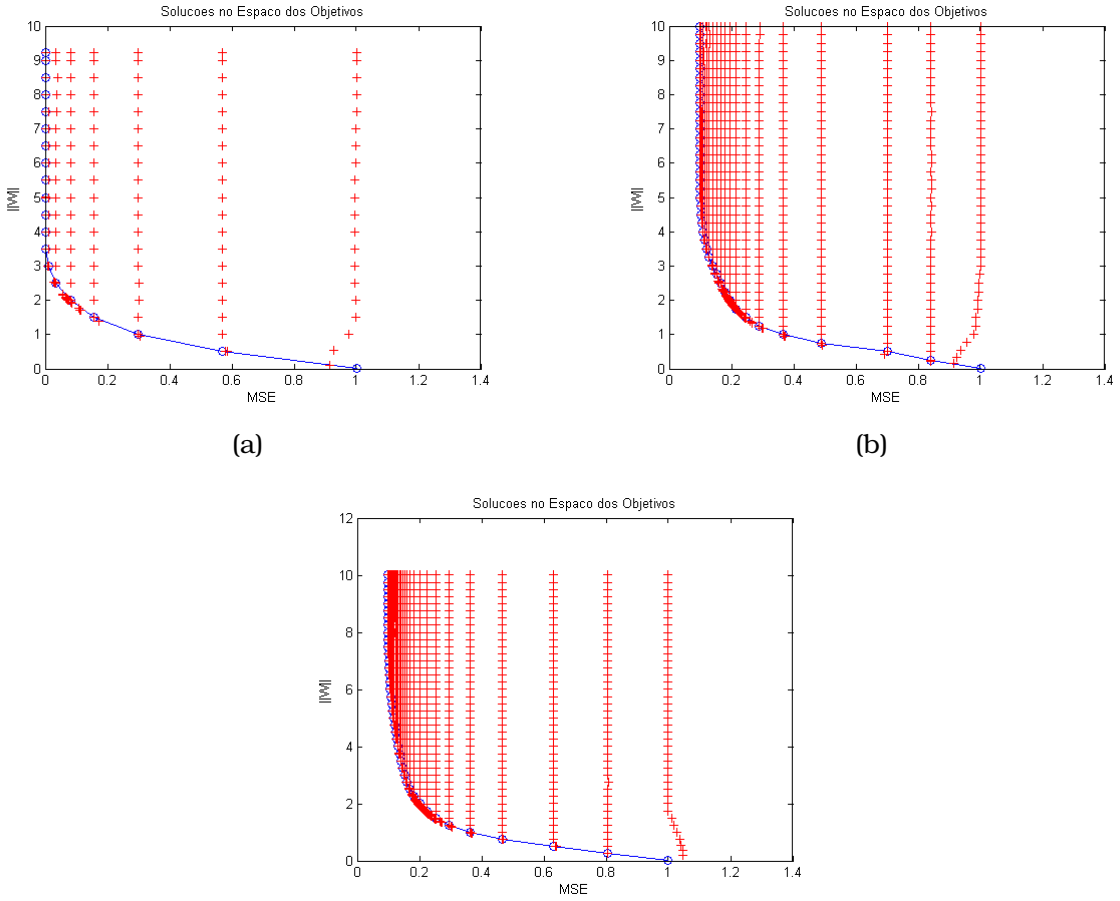


Figura 5.2: Grids de soluções gerados para cada um dos conjuntos de sondas selecionados. (a) conjunto de 30 sondas de Natowicz, (b) conjunto de 18 sondas e (c) conjuntos de 11 sondas.

5.2 A margem geométrica

Uma vez calculado estes conjuntos de soluções para os padrões de treinamento (indutivos), passou-se a determinar quais os padrões limites que deveriam ser considerados no cálculo da margem geométrica tanto do conjunto indutivo como do conjunto de dados transdutivo. Para o conjunto indutivo foi aplicado o método descrito no ítem 3.3 de identificação dos padrões a serem considerados no cálculo da margem geométrica, e para o conjunto de dados transdutivo foram aplicados os métodos MILP e MILFAT descritos no ítem 3.2.

Na seleção destes padrões limites existem alguns parâmetros a serem configurados, e durante as simulações eles foram sistematicamente testados para averiguar qual a sensibilidade de cada um no resultado final. São muitas as combinações possíveis, e nem todas foram testadas, contudo foi possível avaliá-los e mais a frente eles serão discutidos.

Selecionados os padrões limites, calculou-se a margem geométrica das três formas indicadas no ítem 3.4 para cada combinação de métodos. Ou seja,

com os pontos limites selecionados para o conjunto indutivo mais os pontos limites do conjunto transdutivo selecionados pelo método MILP calculou-se a margem geométrica total. Depois, com os mesmos padrões limites selecionados para o conjunto indutivo mais aqueles selecionados pelo método MILFAT para o conjunto transdutivo calculou-se novamente a margem geométrica, e em ambos os casos, calculou-se das três maneiras discutidas, para comparação posterior dos resultados. Todo este procedimento foi repetido para cada um dos conjuntos de sondas selecionados.

5.3 Resultados

Nesta seção são apresentados os melhores resultados alcançados para cada conjunto de sondas para cada diferente combinação de métodos. Em cada uma destas combinações é calculada a margem geométrica total e então seleciona-se a solução que apresentar a maior destas margens. Para esta solução é calculada a matriz de confusão com as classificações corretas e incorretas nas classes PCR e NOPCR para avaliar o desempenho da rede. Os resultados de falsos positivos e falsos negativos são apresentados a seguir.

Nos ítens que se seguem neste capítulo, ρ_{Soma} representa a margem geométrica total calculada através da soma das distâncias dos padrões limites ao hiperplano de separação (equação 3.1), enquanto $\rho_{Multiplicacao}$ representa a margem geométrica total calculada através da multiplicação das distâncias dos padrões limites ao hiperplano de separação (equação 3.7) e $\rho_{Divisao}$ calculada através da divisão das distâncias dos padrões limites ao hiperplano de separação (equação 3.8).

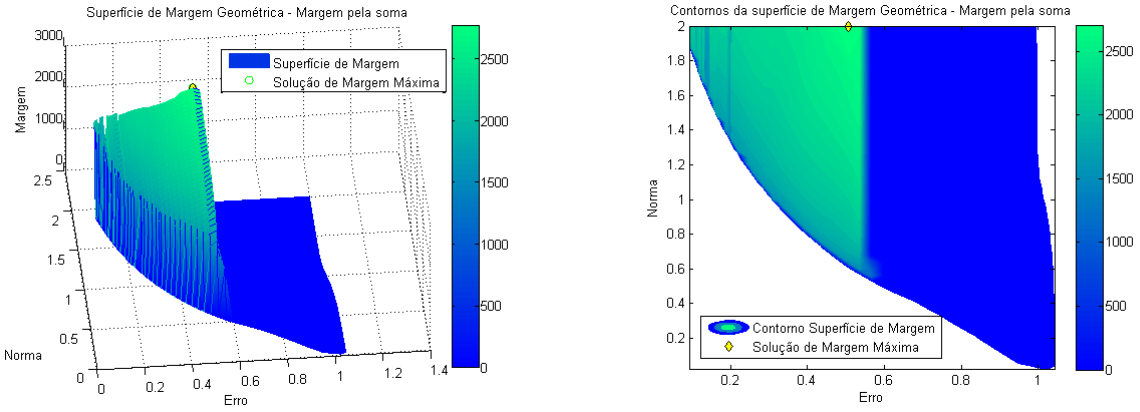
5.3.1 Resultados com o método MILP - 30 sondas

Neste tópico são apresentados os resultados obtidos aplicando o método de identificação de limites de classes rotuladas (MILCL) aos dados indutivos e o método de identificação de limites por probabilidade (MILP) aos dados transdutivos, para o conjunto de 30 sondas. Com os padrões limites selecionados por estes métodos calculou-se a margem geométrica total das três maneiras descritas em 3.4, que aqui serão identificadas como ρ_{Soma}, $\rho_{Multiplicacao}$ e $\rho_{Divisao}$.

A figura 5.3 mostra a superfície de margem geométrica total calculada pela soma e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(a).

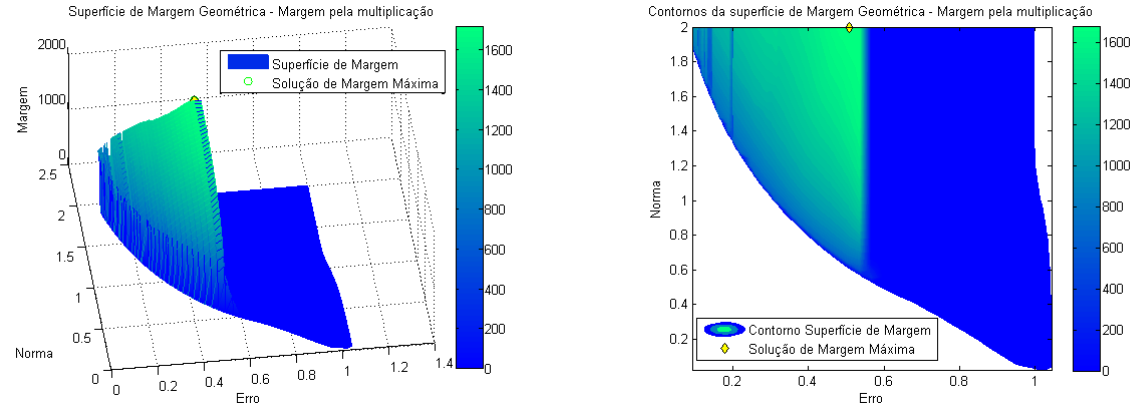
Já a figura 5.4 mostra a superfície de margem geométrica total calculada pela multiplicação e seu contorno no plano Norma x Erro, do mesmo grid.

A figura 5.5 mostra a superfície de margem geométrica total calculada pela



(a) (b)

Figura 5.3: Margem Geométrica Total (soma) - 30 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

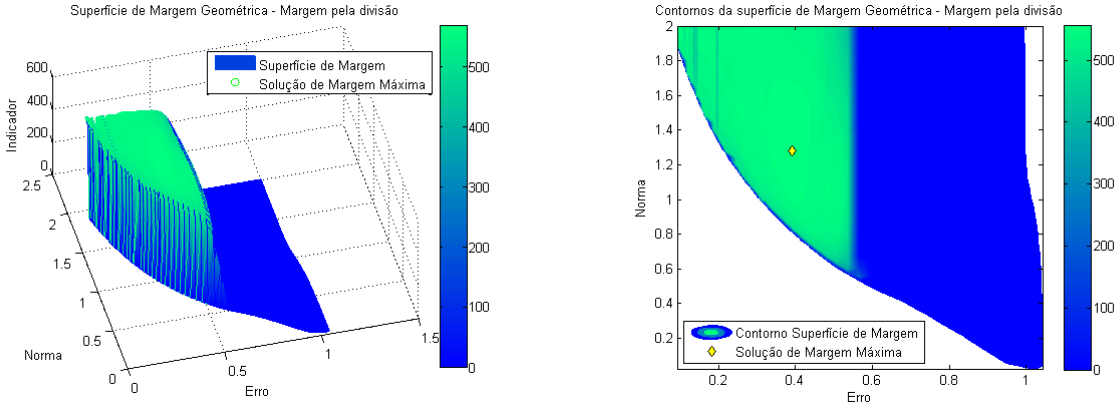


(a) (b)

Figura 5.4: Margem Geométrica Total (multiplicação) - 30 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

divisão e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(a).

Calculou-se então a matriz de confusão para a rede definida pela solução escolhida com o critério da margem geométrica total máxima (ponto amarelo dos gráficos) em cada um dos três diferentes métodos de cálculo da margem. Os resultados são apresentados na tabela 5.1.



(a) (b)

Figura 5.5: Margem Geométrica Total (divisão) - 30 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

Tabela 5.1: Matriz de confusão para os resultados obtidos com o MILP - 30 sondas

	Treinamento		Validação	
Modelo	FP	FN	FP	FN
30 sondas MILP - ρ_{Soma}	11	2	9	1
30 sondas MILP - $\rho_{Multiplicacao}$	11	2	9	1
30 sondas MILP - $\rho_{Divisao}$	9	3	10	1

5.3.2 Resultados com o método MILFAT - 30 sondas

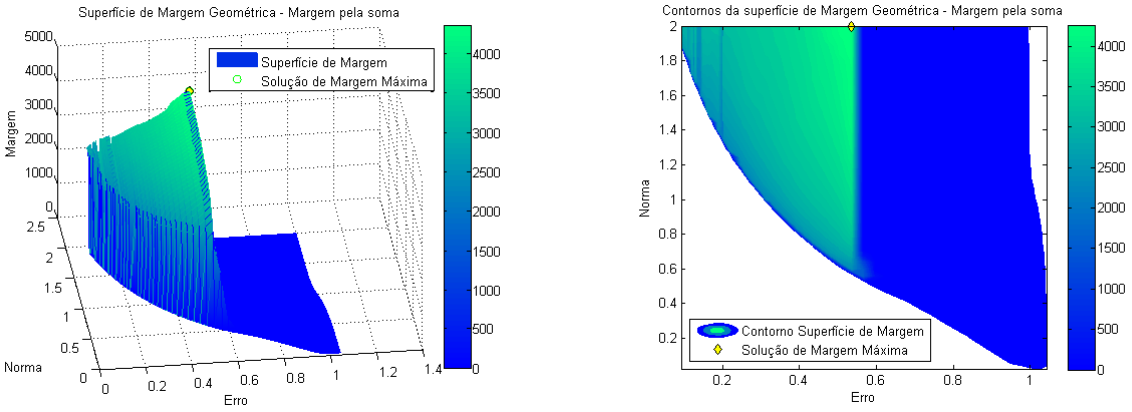
Neste ítem, o que muda em relação ao anterior é o método de seleção dos padrões limites do conjunto transdutivo. Aqui são exibidos os resultados obtidos aplicando o método de identificação de limites de classes rotuladas (MILCL) aos dados indutivos e o método de identificação de limites por fatiamento (MILFAT) aos dados transdutivos, para o conjunto de 30 sondas. Da mesma forma que no ítem anterior, com os padrões limites selecionados por estes métodos calculou-se a margem geométrica total: ρ_{Soma}, $\rho_{Multiplicacao}$ e $\rho_{Divisao}$.

A figura 5.6 mostra a superfície de margem geométrica total calculada pela soma e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(a).

Por sua vez, a figura 5.7 mostra a superfície de margem geométrica total calculada pela multiplicação e seu contorno no plano Norma x Erro, do mesmo grid.

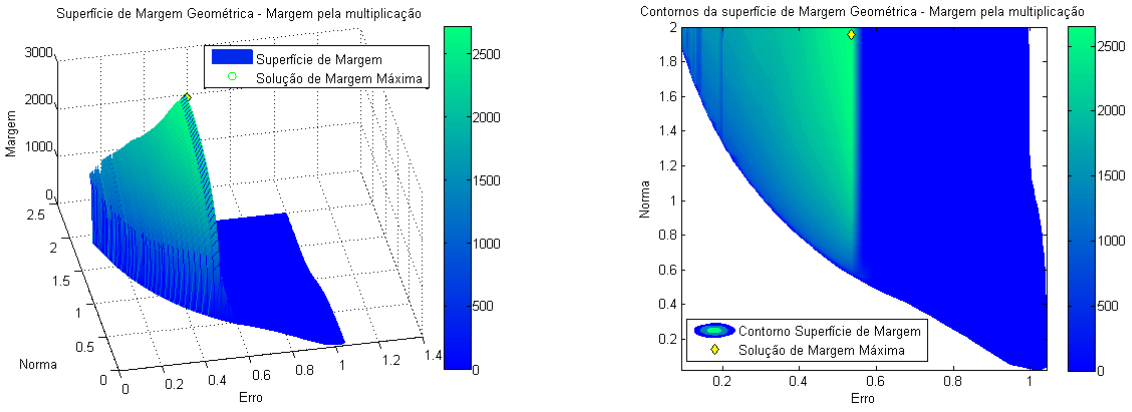
A superfície de margem geométrica total calculada pela divisão e seu contorno no plano Norma x Erro é mostrada na figura 5.8.

A matriz de confusão para a rede definida pela solução escolhida com o



(a) (b)

Figura 5.6: Margem Geométrica Total (soma) - 30 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.



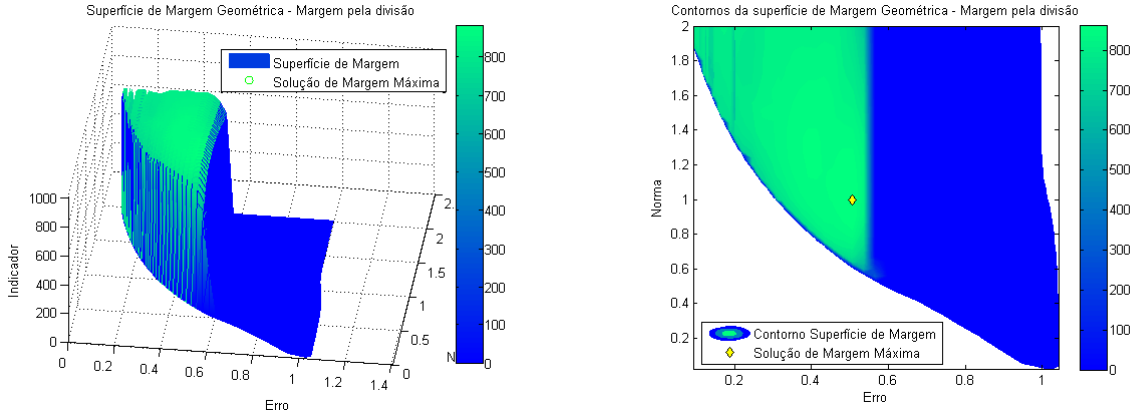
(a) (b)

Figura 5.7: Margem Geométrica Total (multiplicação) - 30 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

critério da margem geométrica total máxima (ponto amarelo dos gráficos) em cada um dos três diferentes métodos de cálculo da margem é apresentada na tabela 5.2.

Tabela 5.2: Matriz de confusão para os resultados obtidos com o MILFAT - 30 sondas

	Treinamento		Validação	
Modelo	FP	FN	FP	FN
30 sondas MILFAT - ρ_{Soma}	13	3	8	1
30 sondas MILFAT - $\rho_{Multiplicacao}$	13	3	8	1
30 sondas MILFAT - $\rho_{Divisao}$	12	3	8	1



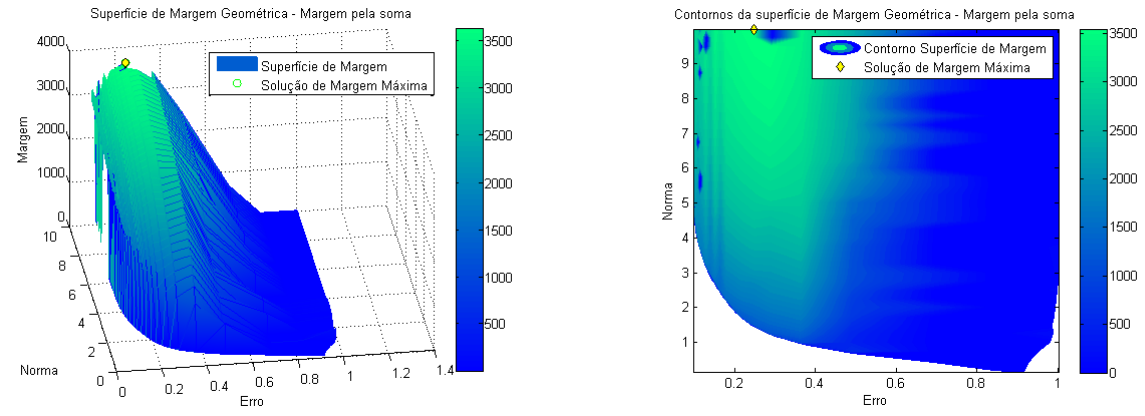
(a) (b)

Figura 5.8: Margem Geométrica Total (divisão) - 30 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

5.3.3 Resultados com o método MILP - 18 sondas

Neste tópico são apresentados os resultados obtidos aplicando o método de identificação de limites de classes rotuladas (MILCL) aos dados indutivos e o método de identificação de limites por probabilidade (MILP) aos dados transdutivos, para o conjunto de 18 sondas. Com os padrões limites selecionados por estes métodos calculou-se a margem geométrica.

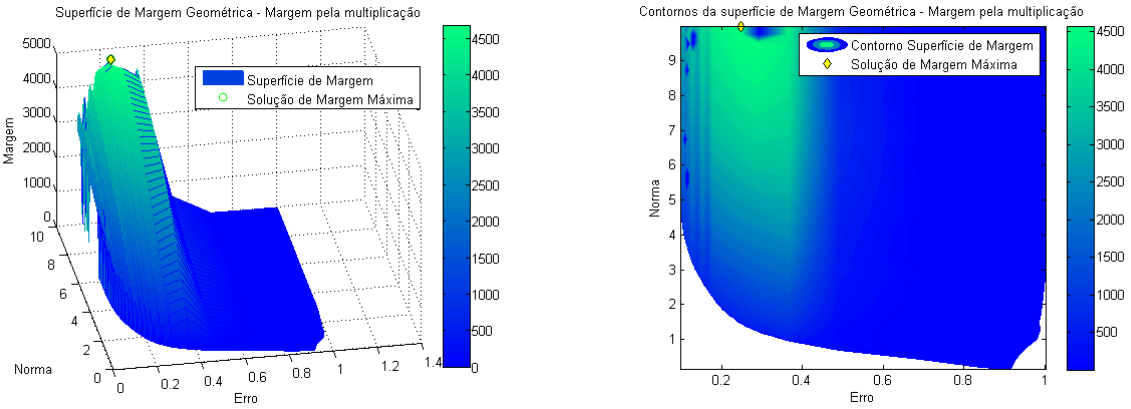
A figura 5.9 mostra a superfície de margem geométrica total calculada pela soma e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(b).



(a) (b)

Figura 5.9: Margem Geométrica Total (soma) - 18 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

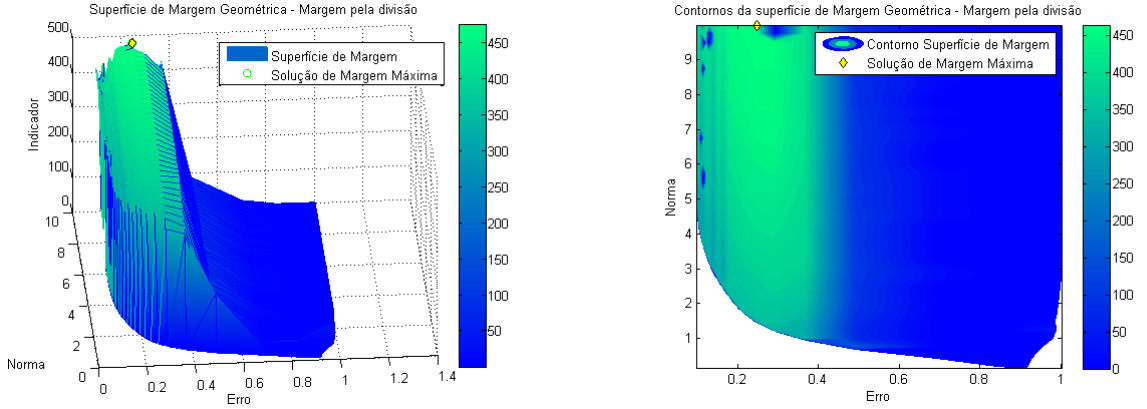
Já a figura 5.10 mostra a superfície de margem geométrica total calculada pela multiplicação e seu contorno no plano Norma x Erro, do mesmo grid.



(a) (b)

Figura 5.10: Margem Geométrica Total (multiplicação) - 18 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

A figura 5.11 mostra a superfície de margem geométrica total calculada pela divisão e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(b).



(a) (b)

Figura 5.11: Margem Geométrica Total (divisão) - 18 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

Calculou-se então a matriz de confusão para a rede definida pela solução escolhida com o critério da margem geométrica total máxima (ponto amarelo dos gráficos) em cada um dos três diferentes métodos de cálculo da margem. Os resultados são apresentados na tabela 5.3.

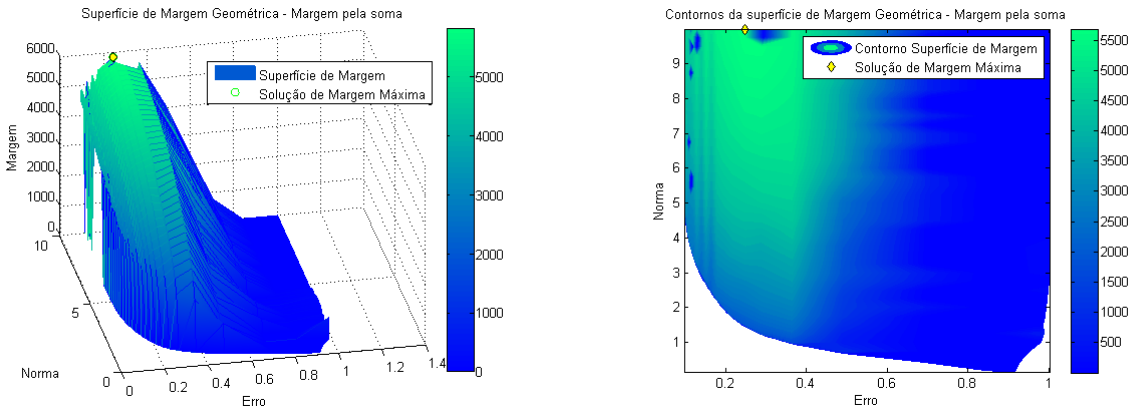
Tabela 5.3: Matriz de confusão para os resultados obtidos com o MILP - 18 sondas

	Treinamento		Validação	
Modelo	FP	FN	FP	FN
18 sondas MILP - ρ_{Soma}	2	3	7	2
18 sondas MILP - $\rho_{Multiplicacao}$	2	3	7	2
18 sondas MILP - $\rho_{Divisao}$	2	3	7	2

5.3.4 Resultados com o método MILFAT - 18 sondas

Neste ítem, o que muda em relação ao anterior é o método de seleção dos padrões limites do conjunto transdutivo. Aqui são exibidos os resultados obtidos aplicando o método de identificação de limites de classes rotuladas (MILCL) aos dados indutivos e o método de identificação de limites por fatiamento (MILFAT) aos dados transdutivos, para o conjunto de 18 sondas. Da mesma forma que no ítem anterior, com os padrões limites selecionados por estes métodos calculou-se a margem geométrica total: ρ_{Soma}, $\rho_{Multiplicacao}$ e $\rho_{Divisao}$.

A figura 5.12 mostra a superfície de margem geométrica total calculada pela soma e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(b).

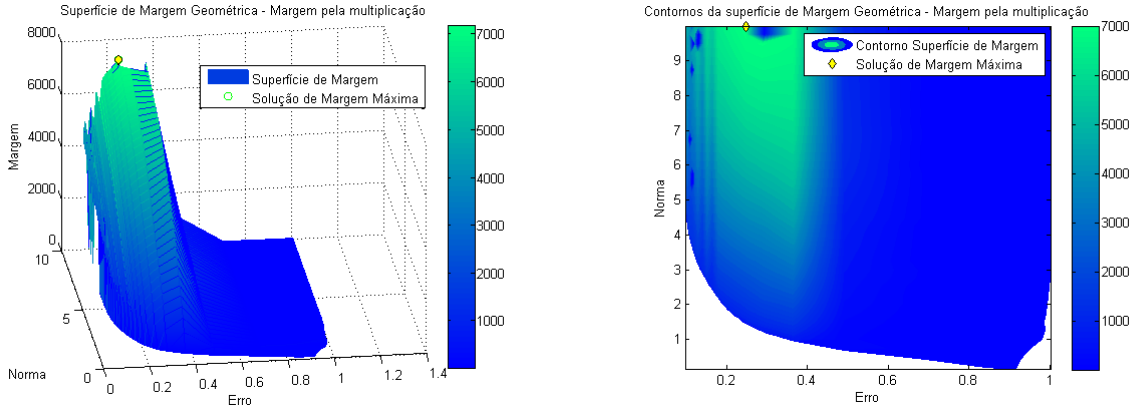


(a) (b)

Figura 5.12: Margem Geométrica Total (soma) - 18 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

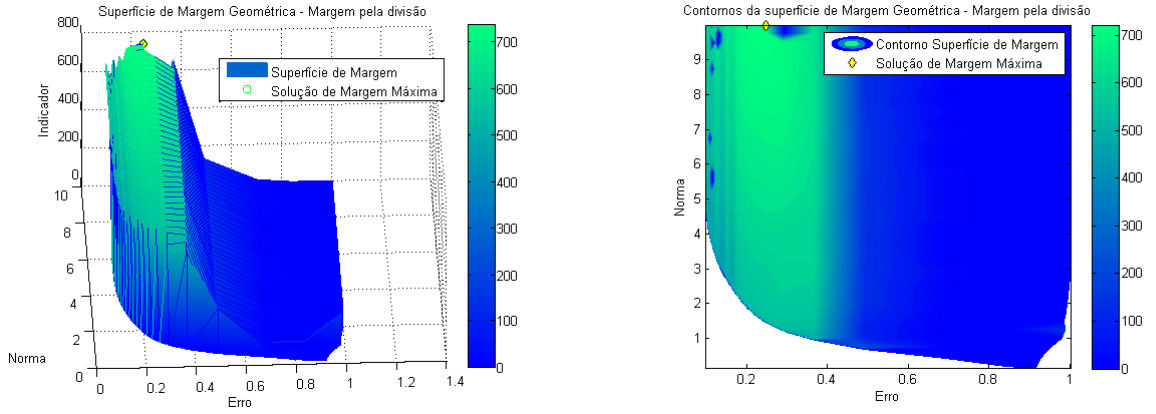
Por sua vez, a figura 5.13 mostra a superfície de margem geométrica total calculada pela multiplicação e seu contorno no plano Norma x Erro, do mesmo grid.

A superfície de margem geométrica total calculada pela divisão e seu contorno no plano Norma x Erro é mostrada na figura 5.14.



(a) (b)

Figura 5.13: Margem Geométrica Total (multiplicação) - 18 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.



(a) (b)

Figura 5.14: Margem Geométrica Total (divisão) - 18 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

A matriz de confusão para a rede definida pela solução escolhida com o critério da margem geométrica total máxima (ponto amarelo dos gráficos) em cada um dos três diferentes métodos de cálculo da margem é apresentada na tabela 5.4.

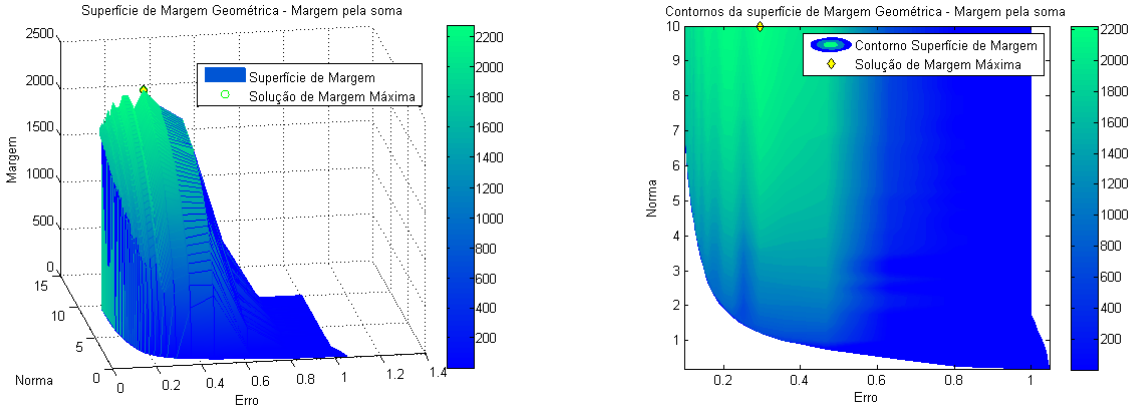
Tabela 5.4: Matriz de confusão para os resultados obtidos com o MILFAT - 18 sondas

	Treinamento		Validação	
Modelo	FP	FN	FP	FN
18 sondas MILFAT - ρ_{Soma}	2	3	7	2
18 sondas MILFAT - $\rho_{Multiplicao}$	2	3	7	2
18 sondas MILFAT - ρ_{Diviso}	2	3	7	2

5.3.5 Resultados com o método MILP - 11 sondas

Neste tópico são apresentados os resultados obtidos aplicando o método de identificação de limites de classes rotuladas (MILCL) aos dados indutivos e o método de identificação de limites por probabilidade (MILP) aos dados transdutivos, para o conjunto de 11 sondas. Com os padrões limites selecionados por estes métodos calculou-se a margem geométrica.

A figura 5.15 mostra a superfície de margem geométrica total calculada pela soma e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(c).



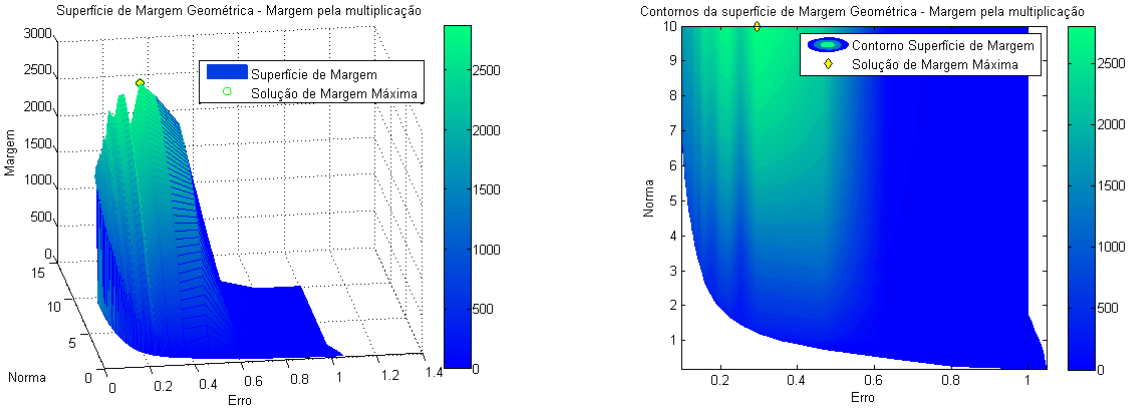
(a) (b)

Figura 5.15: Margem Geométrica Total (soma) - 11 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

Por sua vez, a figura 5.10 mostra a superfície de margem geométrica total calculada pela multiplicação e seu contorno no plano Norma x Erro, do mesmo grid.

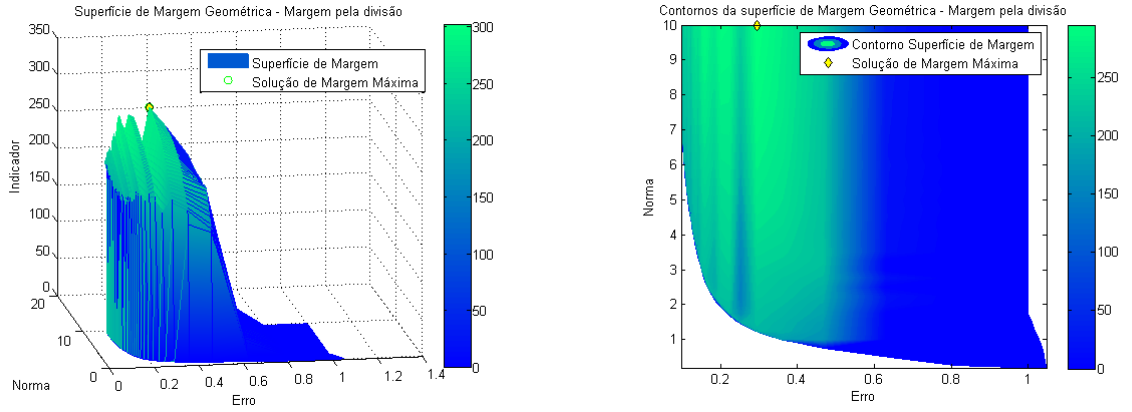
A figura 5.17 mostra a superfície de margem geométrica total calculada pela divisão e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(c).

Foi calculada então a matriz de confusão para a rede definida pela solução escolhida com o critério da margem geométrica total máxima (ponto amarelo dos gráficos) em cada um dos três diferentes métodos de cálculo da margem. Os resultados são apresentados na tabela 5.5.



(a) (b)

Figura 5.16: Margem Geométrica Total (multiplicação) - 11 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.



(a) (b)

Figura 5.17: Margem Geométrica Total (divisão) - 11 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

Tabela 5.5: Matriz de confusão para os resultados obtidos com o MILP - 11 sondas

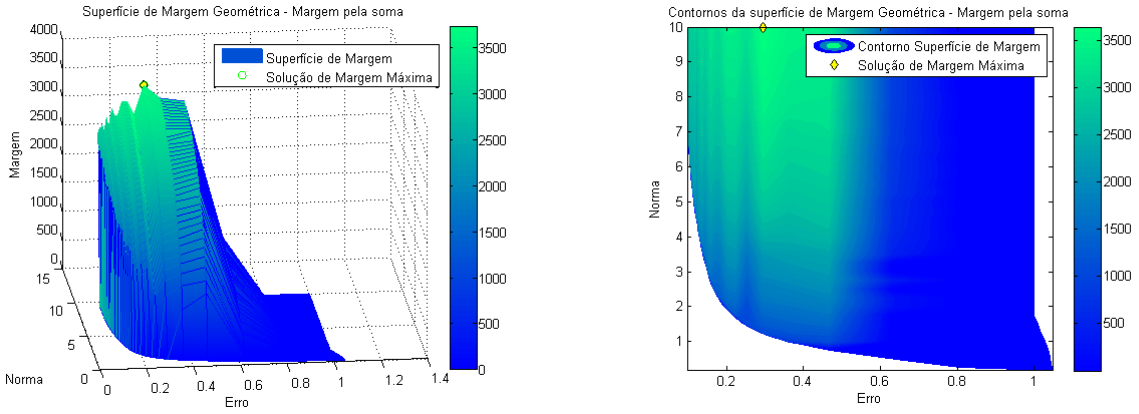
	Treinamento		Validação	
Modelo	FP	FN	FP	FN
11 sondas MILP - ρ_{Soma}	3	3	8	2
11 sondas MILP - $\rho_{Multiplicacao}$	3	3	8	2
11 sondas MILP - $\rho_{Divisao}$	3	3	8	2

5.3.6 Resultados com o método MILFAT - 11 sondas

Da mesma forma que nos ítens anteriores, neste ítem, o que muda em relação ao anterior é o método de seleção dos padrões limites do conjunto transdutivo. Aqui são exibidos os resultados obtidos aplicando o método de

identificação de limites de classes rotuladas (MILCL) aos dados indutivos e o método de identificação de limites por fatiamento (MILFAT) aos dados transdutivos, para o conjunto de 11 sondas. Da mesma forma que no ítem anterior, com os padrões limites selecionados por estes métodos calculou-se a margem geométrica total: ρ_{Soma}, $\rho_{Multiplicacão}$ e $\rho_{Divisão}$.

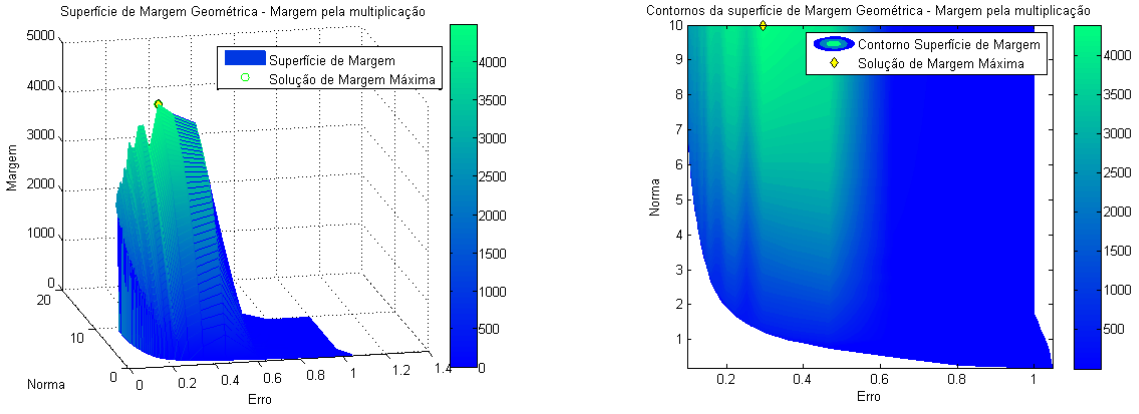
A figura 5.18 mostra a superfície de margem geométrica total calculada pela soma e seu contorno no plano Norma x Erro, do grid apresentado na figura 5.2(c).



(a) (b)

Figura 5.18: Margem Geométrica Total (soma) - 11 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

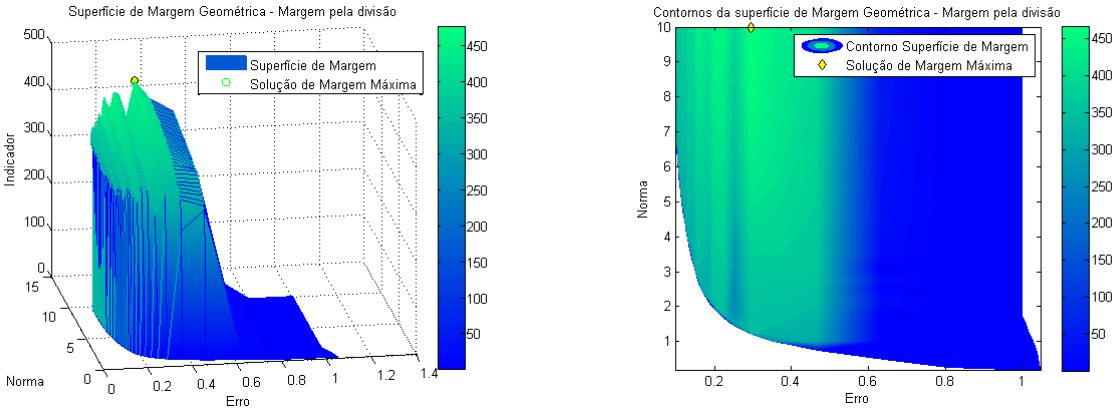
Por sua vez, a figura 5.19 mostra a superfície de margem geométrica total calculada pela multiplicação e seu contorno no plano Norma x Erro, do mesmo grid.



(a) (b)

Figura 5.19: Margem Geométrica Total (multiplicação) - 11 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

A superfície de margem geométrica total calculada pela divisão e seu contorno no plano Norma x Erro é mostrada na figura 5.20.



(a) (b)

Figura 5.20: Margem Geométrica Total (divisão) - 11 sondas . (a) Superfície e (b) os contornos da Margem Geométrica. O ponto amarelo indica a solução de margem total máxima.

A matriz de confusão para a rede definida pela solução escolhida com o critério da margem geométrica total máxima (ponto amarelo dos gráficos) em cada um dos três diferentes métodos de cálculo da margem é apresentada na tabela 5.6.

Tabela 5.6: Matriz de confusão para os resultados obtidos com o MILFAT - 11 sondas

	Treinamento		Validação	
Modelo	FP	FN	FP	FN
11 sondas MILFAT - ρ_{Soma}	3	3	8	2
11 sondas MILFAT - $\rho_{Multiplicacão}$	3	3	8	2
11 sondas MILFAT - $\rho_{Divisao}$	3	3	8	2

5.4 Considerações finais

Os resultados encontrados pelo método semi-supervisionado de aprendizagem baseado na margem geométrica são muito bons e promissores. Os dados de entrada são esparsos, reduzidos, de alta dimensão e sem uma região de separação bem definida, no entanto o método convergiu e conseguiu um boa generalização. No capítulo 6 faz-se uma discussão completa dos resultados, mas aqui apresenta-se uma tabela resumo com os resultados alcançados e uma segunda tabela com os resultados encontrados por Horta em seu trabalho [36] para efeito de comparação.

Os melhores resultados do método de apredizado semi-supervisionado proposto ocorreram para o conjunto de 18 sondas. Este resultado é muito semelhante aos obtidos por Horta com os modelos *18 sondas SVM RBF*, *32 sondas SVM RBF* e *18 sondas LASSO* nas três primeiras linhas da tabela 5.8. Os resultados obtidos com o conjunto de 11 sondas também são comparáveis aos de Horta citados anteriormente, mas são pouco "piores" que os obtidos com 18 sondas. Já os obtidos com o conjunto de 30 sondas são comparáveis com os piores resultados obtidos por Horta e mostrados nas linhas finais da tabela 5.8. As demais discussões que se fazem necessárias são discutidas no capítulo 6.

Tabela 5.7: Matriz de confusão para os resultados obtidos com MSMG

Modelo	Treinamento		Validação	
	FP	FN	FP	FN
18 sondas MILP - ρ_{Soma}	2	3	7	2
18 sondas MILP - $\rho_{Multiplicacao}$	2	3	7	2
18 sondas MILP - $\rho_{Divisao}$	2	3	7	2
18 sondas MILFAT - ρ_{Soma}	2	3	7	2
18 sondas MILFAT - $\rho_{Multiplicao}$	2	3	7	2
18 sondas MILFAT - ρ_{Diviso}	2	3	7	2
11 sondas MILP - ρ_{Soma}	3	3	8	2
11 sondas MILP - $\rho_{Multiplicacao}$	3	3	8	2
11 sondas MILP - $\rho_{Divisao}$	3	3	8	2
11 sondas MILFAT - ρ_{Soma}	3	3	8	2
11 sondas MILFAT - $\rho_{Multiplicacão}$	3	3	8	2
11 sondas MILFAT - $\rho_{Divisao}$	3	3	8	2
30 sondas MILP - ρ_{Soma}	11	2	9	1
30 sondas MILP - $\rho_{Multiplicacao}$	11	2	9	1
30 sondas MILP - $\rho_{Divisao}$	9	3	10	1
30 sondas MILFAT - ρ_{Soma}	13	3	8	1
30 sondas MILFAT - $\rho_{Multiplicacao}$	13	3	8	1
30 sondas MILFAT - $\rho_{Divisao}$	12	3	8	1

Tabela 5.8: Matriz de confusão para os resultados obtidos por Horta [36]

Modelo	Treinamento		Validação	
	FP	FN	FP	FN
18 sondas SVM RBF	1	3	5	2
32 sondas SVM RBF	2	3	6	1
18 sondas LASSO	2	4	4	3
18 sondas Naïve Bayes	6	4	5	2
18 sondas SVM Linear	9	4	5	2
18 sondas MLP	3	5	6	2
10 sondas SVM Linear	7	3	6	2
Natowicz Voto Majoritário	9	4	6	1
10 sondas SVM RBF	9	4	6	2
Natowicz DLDA	11	3	7	2
Hess DLDA	14	1	11	1

CAPÍTULO 6

Discussões Conclusões

Inicialmente este capítulo discutirá os resultados obtidos no trabalho e que foram apresentados no capítulo 5. O método alcançou resultados tão melhores quanto melhores foram seus dados de entrada, ou quanto mais significativos foram. Os piores resultados foram obtidos com o conjunto de 30 sondas. Talvez o excesso de informações acabou por "mascarar" a região de separação das classes já que o número de dados para cada padrão de entrada é quase da mesma ordem de grandeza que o número de padrões, dificultando muito alcançar bons resultados.

Os resultados obtidos com 18 e 11 sondas são comparáveis aos melhores resultados alcançados por Horta em seu trabalho [36]. São resultados equilibrados, ou seja, classificam bem tanto o conjunto de treinamento quanto o de teste. Se for feita uma análise como a proposta em [44], onde plota-se um gráfico da distância da solução, para o conjunto de treinamento, ao ponto ótimo (0,1) do espaço ROC, pela mesma distância só que para o conjunto de teste, teremos um ponto bem próximo à reta que indicando a estabilidade das soluções. Esta maneira de visualizar a solução permite se ter uma noção melhor da performance de suas sensibilidade e especificidade. Quanto mais próximo da reta definida pela distância de teste igual à distância de treinamento mais estável é a solução, e quanto mais próximo à origem deste gráfico menos erros são cometidos.

O método semi-supervisionado baseado na margem geométrica aplicado ao problema de previsão da eficácia da quimeoterapia neoadjuvante no tratamento de câncer de mama apresentou resultados promissores, mas faz-se necessário discutir alguns pontos. O primeiro ponto é com relação ao cálculo da margem geométrica. Apesar de que, para o problema em questão, os re-

sultados dos diferentes cálculos de margem deram resultados muito próximos para um mesmo conjunto de sondas, a discussão de como calculá-la procede e é de suma importância. Realmente tem-se então a impressão de que foi em vão tanta discussão a este respeito, contudo, este resultado pode estar intimamente ligado à natureza do problema e sobretudo ao reduzido número de amostras.

Outro ponto a se ressaltar é quanto aos parâmetros que devem ser configurados ao longo do processo. O método de identificação de limites por fatiamento envolve ter de definir um número de faixas para a pesquisa, e, como o algoritmo tem um elevado custo computacional este fator é de extrema importância: quanto mais faixas, mais demorado ele será. Além disso, um grande número de fatias pode causar a seleção de padrões mais internos às classes. Já no método de identificação de limites por probabilidade, tem-se que definir um ponto de corte da probabilidade de um padrão pertencer ou não ao limite da partição. E mais uma vez surge a questão: padrões com probabilidades inferiores a que número podem ser considerados como limites? O número de vezes em que se vai executar o FCM em ambos os métodos e a quantidade de partições em cada uma delas também pode influenciar o resultado, contudo estes parâmetros são de menor importância. Foram feitas diversas simulações alterando estes valores e a grande maioria das vezes chegavam-se aos mesmos resultados. Apenas quando se configurava números de partições muito baixos ou extremamente elevados acabavam distorcendo os resultados. O último parâmetro que se quer discorrer é sobre o número de vezes em que cada padrão foi considerado como limite para a seleção final. Com as simulações percebeu-se que deve-se dar preferência aos padrões selecionados em todas as iterações do FCM. Outra constatação é que estes métodos podem selecionar todos os padrões de entrada como sendo limites, para problemas muito esparsos, o que pode não ser mesmo um problema, porém esta consideração não foi exaustivamente verificada.

O último ponto de discussão é também relacionado à seleção dos pontos limites, contudo, relacionado ao conjunto de dados a ser utilizado neste processo. Ao tentar extrair a informação de margem geométrica do conjunto transdutivo procura-se determinar os padrões limites que separam as classes em potencial. Espera-se que entre estas classes exista uma região de baixa densidade em relação à quantidade de padrões. Já para o conjunto indutivo é bem mais simples: não é necessário artifícios para gerar uma "classificação" para então determinar os limites, pois já se conhece os rótulos destes padrões. Mas, os padrões do conjunto indutivo também fazem parte do problema assim como sua distribuição. Neste trabalho a identificação dos limites foi feita separadamente para cada um destes conjuntos, todavia, separar o conjunto

indutivo do transdutivo ao se tentar identificar as regiões de baixa densidade pode inserir uma distorção ao resultado, uma vez que a ausência dos padrões de treinamento no momento do processamento vai ser traduzido como uma região de baixa densidade. Basta analisar o problema das duas luas para se entender o que está-se discutindo aqui. O conjunto transdutivo sem o indutivo apresenta uma área de baixa densidade diferente e mais ampla que a que realmente existe. Assim, se ao lidar com o conjunto transdutivo os padrões de treinamento forem considerados, mesmo sem a informação dos rótulos, pode-se contribuir para a determinação mais precisa dos limites.

O principal objetivo deste trabalho era estudar e discutir formas de se calcular a margem geometricamente. Todo problema de classificação, mesmo com superposição, deve ter uma separação e, nesta região que encerra esta separação, existe um hiperplano ótimo de separação. Uma vez que se conhece esta superfície então é possível calcular sua margem geometricamente. Iniciou-se desta forma uma intensa fase de simulações e cálculos tentando esgotar ao máximo todos os ângulos do problema. Todas as linhas de estudo adotadas, ao longo do tempo, serviram em sua grande maioria, apenas, para trazer mais uma perspectiva da grande complexidade e do grande desafio deste problema. Ao final consegiu-se chegar a uma forma, ou pelo menos a uma direção, onde é possível extrair uma informação geométrica dos dados e utilizá-la para definir se uma solução é melhor que outra para a resolução do problema.

Para calcular a margem geométrica de um conjunto de dados não classificados (*unlabeled*) foi preciso resolver um grande problema em função de uma decisão tomada. Decidiu-se tentar identificar as regiões de baixa densidade entre as classes, assumindo que elas existam e que nelas esteja contido o hiperplano ótimo. Foi esta a abordagem principal deste trabalho. Diversas tentativas e experimentos foram feitos na intenção de se determinar estas regiões o que levou aos métodos descritos neste trabalho, mas com um elevado custo computacional, o que não o invalida. Apesar de ter um alto custo computacional, ele pode ser aplicado em trabalhos acadêmicos para resolver problemas inclusive de dimensões mais elevadas e seus resultados continuarão a ser válidos.

Ao aplicar este método semi-supervisionado ao problema do tratamento quimeoterápico pode-se perceber como é importante a identificação e seleção das informações que são mais relevantes para sua solução. Os resultados obtidos para um conjunto de informações maior para cada padrão de entrada, foram piores do que se esperava, afinal, mais informação pode soar bem aos ouvidos. Mas mais informação simplesmente não ajuda, mas mais informação adequada é a grande questão. Mais sondas por paciente ao invés de ajudar

atrapalhou, pois como são poucos os clientes, estas informações acabaram por tomar a mesma ordem de grandeza e congestionaram a rede. Tanto é verdade que ao reduzir o conjunto de sondas por paciente o resultado melhorou consideravelmente. Também é verdade, que ao reduzir ainda mais o conjunto de informações o resultado, apesar de ainda muito bom, se tornou um pouco pior. Aí já foi informação valiosa de menos. É muito importante ter acesso às informações das 22000 sondas, contudo, selecionar aquelas informações vitais é ainda mais importante. Claro que quanto mais pacientes forem incluídos na base de dados, melhores serão os resultados, mesmo para o conjunto de 30 sondas.

Como propostas para trabalhos futuro deixa-se algumas sugestões. Existem vários pontos abertos e passíveis de discussão: será que existe outra forma mais objetiva e menos custosa computacionalmente para se determinar as áreas de baixa densidade de padrões, e com isso definir mais facilmente e rapidamente os padrões limites? Outra idéia muito interessante é definir uma função de custo baseada nestes princípios explicados que possa ser minimizada em um método multi-objetivo levando a uma solução de margem geométrica máxima que classifique bem o conjunto de treinamento e que tenha boa generalização. Neste capítulo discutiu-se a questão de utilizar o conjunto indutivo, sem a informação dos rótulos, em conjunto com transdutivo para melhor caracterizar as áreas de baixa densidade de padrões. Mas se conhece a classificação dos padrões de treinamento, então, porque não tentar utilizar esta informação na identificação das áreas de baixa densidade, principalmente auxiliando o FCM ou outro método qualquer de agrupamento.

APÊNDICE

A

O problema de aplicação

As células do corpo humano crescem e morrem de forma ordenada e quando isso não ocorre surgem os tumores que podem ser benignos ou malígnos (câncer). Se eles ñão invadem os tecidos próximos ou qualquer outra parte do corpo então são classificados como benignos e não representam risco ao ser humano, caso contrário, se ele se espalha causando destruição dos tecidos, então é classificado como tumor malígno ou câncer e este sim pode levar o paciente à morte.

O câncer de mama é o que mais atinge as mulheres no mundo e seu tratamento envolve três fases distintas:

- Quimeoterapia neoadjuvante, que nada mais é que um tratamento quimeoterápico que tem como intenção evitar a cirurgia ou mesmo reduzir a extensão do tumor.
- Cirurgia, que como o próprio nome sugere, é a intervenção para tentar remover as células cancerígenas.
- Quimeoterapia adjuvante, que é o tratamento quimeoterápico após a cirurgia para evitar a metástase (espalhamento pelo corpo) das células cancerígenas.

Acontece que cerca de apenas 30% dos pacientes respondem bem à quimeoterapia neoadjuvante, e nos outros 70% dos casos, o paciente é submetido a todo o desgaste e sofrimento deste tratamento em vão. Seria muito interessante que os médicos tenham uma maneira de prever se este tratamento seria eficaz ou não em um determinado paciente, evitando assim, sofrimentos desnecessários. Apenas informações clínicas não são capazes ainda de dar

bons resultados neste sentido, contudo, com o surgimento da tecnologia de microarrays abre-se uma nova porta para se chegar a este sonho.

A.1 Composição da mama

As mamas são glândulas cuja função principal é a produção de leite [1]. São compostas de lobos que se dividem em porções menores, os lóbulos, e ductos, que conduzem o leite produzido para fora pelo mamilo. Também existem vasos sanguíneos em toda sua extensão e vasos linfáticos que têm a função de carregar nutrientes e retirar o que não é desejado. Estes vasos linfáticos se agrupam nos gânglios linfáticos e fazem a drenagem destes gânglios da mama para os que existem no tórax, no pescoço ou nas axilas.

A.2 O câncer de mama

Quando as células deste órgão passam a se dividir e a se reproduzir muito rápido e de forma desordenada ocorre o câncer de mama. Os ductos das mamas são os mais acometidos por células cancerígenas e, por isso, a forma mais comum é chamada Carcinoma Ductal. Este câncer pode ser classificado como *in situ*, quando não passa das primeiras camadas de célula destes ductos, ou invasor, quando invade os tecidos em volta. Os que começam nos lóbulos da mama são chamados de Carcinoma Lobular e são menos comuns que o primeiro, e, muito freqüentemente acomete as duas mamas. O Carcinoma Inflamatório de mama é um câncer mais raro e normalmente se apresenta de forma agressiva, comprometendo toda a mama, deixando-a vermelha, inchada e quente.

Existem fatores de risco conhecidos. Alguns destes fatores são modificáveis, ou seja, pode-se alterar a exposição que uma pessoa tem a este determinado fator, diminuindo a sua chance de desenvolver este câncer. Da mesma forma existem fatores de proteção que diminuem as chances de desenvolvê-los.

São os fatores de risco e proteção:

- Idade
- Exposição excessiva a hormônios
- Radiação
- Dieta
- Exercício físico
- História ginecológica

- História familiar
- Alterações nas mamas

A.3 Sintomas do câncer de mama

O primeiro fato importante é que o câncer de mama normalmente não dói. Apenas pode-se sentir a presença de um nódulo (ou caroço) que anteriormente não era sentido. Ao se detectar um nódulo a pessoa deve procurar um médico para uma melhor avaliação através de exame clínico e uma mamografia.

Outro sintoma é uma deformidade na suas mamas, ou as mamas podem estar assimétricas, ou ainda pode-se notar uma retração na pele ou um líquido sanguinolento saindo pelo mamilo. Nos casos mais adiantados pode aparecer uma ferida (ulceração) na pele com odor muito desagradável [1].

No caso de carcinoma inflamatório a mama pode aumentar rapidamente de volume, ficando quente e vermelha.

A.4 Tratamento

Como mencionado na introdução deste capítulo o tratamento do câncer de mama pode ter até três fases ou etapas descritas a seguir.

Quimeoterapia Neoadjuvante

É aplicada geralmente a pacientes com câncer localmente avançado e tem por objetivo tentar diminuir ou mesmo extinguir o tumor antes da cirurgia, de forma que a mama possa ser parcial ou totalmente preservada.

Cirurgia

Existem dois tipos de procedimentos que podem ser adotados dependendo do tumor da paciente. O primeiro tipo é chamado cirurgia conservadora e é indicada para pacientes que tem partes da mama livres da doença. Já o segundo tipo é a mastectomia onde toda a mama é retirada sendo uma cirurgia mais radical e que causa grandes impactos psicológicos nas pacientes.

Quimeoterapia Adjuvante

Tem por objetivo evitar o reaparecimento do tumor ou mesmo a metástase que é o espalhamento do câncer por todo o corpo. É aplicado após a cirurgia.

A.5 Considerações finais

O problema do tratamento do câncer de mama não está restrito a um grupo específico de mulheres mas alcança mulheres (e até homens o que é mais raro) no mundo inteiro. Deve este problema ser tratado então como um problema de saúde pública. Realmente, estimular a prevenção diminui os riscos de um tratamento mais longo e doloroso quando diagnosticado positivamente, e até mesmo reduz os custos com este tratamento.

Uma vez diagnosticado positivamente, poder prever se o tratamento quimeoterápico pré-cirúrgico será eficaz ou não para um determinado paciente tem as mesmas vantagens: reduz o sofrimento destas pessoas que não vão passar por ele em vão e até reduz os custos com o tratamento.

Referências Bibliográficas

[1] http://www.abcdasaude.com.br/artigo.php?611.

[2] V. Ben-Hur, A.; Vapnik. A suport vector method for clustering.

[3] C. Bennett, K.; Campbell. Support vector machines: Hype or hallelujah? *SIGKDD Explorations.*, 2000.

[4] J. Bennett, K.; Bi. A geometric approach to support vector regression. *Elsevier Science*, May 2003.

[5] K. Bennett. Semi-supervised suport vector machines.

[6] Kristin P. Bennett and Erin J. Bredensteiner. Geometry in learning. In *Geometry at Work*, 1997.

[7] Kristin P. Bennett, Nello Cristianini, John Shawe-taylor, and Donghui Wu. Enlarging the margins in perceptron decision trees. In *Machine Learning*, pages 295–313, 2000.

[8] J.C. Bezdec. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[9] Jinbo Bi and Kristin P. Bennett. Duality, geometry, and support vector regression. In *Advances in Neural Information Processing Systems*, pages 593–600. MIT Press, 2002.

[10] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] Antônio P. Braga, Euler G. Horta, René Natowicz, Roman Rouzier, Roberto Incitti, Thiago S. Rodrigues, Marcelo A. Costa, Carmen D. M. Pataro, and Arben Çela. Bayesian classifiers for predicting the outcome of breast cancer preoperative chemotherapy. *In The Third International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR08)*.

[12] A.P. Braga. *Redes Neurais Artificiais. Teoria e aplicações.* LTC Editora, Rio de Janeiro, 2000.

[13] C. Bugres. A tutorial on support vector machines on pattern recognition. *Kluwer Academic publishers, Boston.*, 1998.

[14] B. Carvalho. O estado da arte em métodos para reconhecimento de padrões: Support vector machine. *SUCESU*, 2005.

[15] Jesús Cid-Sueiro and José L. Sancho-Gómez. Saturated perceptrons for maximum margin and minimum misclassification error. *Neural Process. Lett.*, 14(3):217–226, 2001.

[16] Marcelo Azevedo Costa, Antônio de Pádua Braga, and Benjamin Rodrigues de Menezes. Improving generalization of mlps with sliding mode control and the levenberg-marquardt algorithm. *Neurocomput.*, 70(7-9):1342–1347, 2007.

[17] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in patter recognition. *IEEE Transactions on Eletronic Computers, vol. EC-14,pp.326-334*, 1965.

[18] K. Crammer, R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin analysis of the lvq algorithm, 2002.

[19] D. Crisp. A geometric interpretation of v-svm classifiers.

[20] Xiangqin Cui and Gary Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome Biology, 4(4):210*, 2003.

[21] A.; Bartlett P.;Scholkopf B. Dale, S.; Smola. *Advances in Large Margin Classifiers*. The MIT Press, 2000.

[22] Carlotta Domeniconi and Dimitrios Gunopulos. Adaptive nearest neighbor classification using support vector machines. In *In Advances in Neural Information Processing Systems 14*, pages 665–672. MIT Press, 2002.

[23] D.; Peng J. Domeniconi, C.; Gunopulos. Large margin nearest neighbor classifiers. *IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 4*, JULY 2005.

[24] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[25] S.E. Fahlman. An empirical study of learning speed in back-propagation networks. *Technical Report, Carnegie Mellow University*, 1988.

[26] P. A. V. Ferreira. Otimização multiobjetivo: Teoria e aplicações. tese de livre docência., 1999.

[27] A.K. JAIN; M.N. MURTY; P.J. FLYNN. Data clustering: A review. *ACM Computing Surveys, Vol. 31, No. 3*, September 1999.

[28] Vojtech Franc and Václav Hlavác. An iterative algorithm learning the maximal margin classifier. *Pattern Recognition*, 36(9):1985 – 1996, 2003. Kernel and Subspace Methods for Computer Vision.

[29] Y Freund and R. Schapire. Large margin classification using the perceptron algorithm. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

[30] Claudio Gentile. A new approximate maximal margin classification algorithm. *J. Mach. Learn. Res.*, 2:213–242, 2002.

[31] Tony Van Gestel, Johan A. K. Suykens, Bart Baesens, Stijn Viaene, Jan Vanthienen, Guido Dedene, Bart De Moor, and Joos Vandewalle. Benchmarking least squares support vector machine classifiers. *Mach. Learn.*, 54(1):5–32, 2004.

[32] S. Gunn. Support vector machines for classification and regression. *Image Speech & Intelligent Systems Group - University of southampton*, 1997.

[33] M. Hagan and M. Menhaj. *Training feedforward networks with the Marquardt algorithm.* IEEE Transactions on Newral Networks, 5(6):989-993, November 1994.

[34] S. Haykin. *Redes Neurais: Princípios e Prática*. Bookman, 2001.

[35] KR Hess, K Anderson, WF Symmans, V Valero, N Ibrahim, JA Mejia, D Booser, RL Theriault, AU Buzdar, PJ Dempsey, R Rouzier, N Sneige, JS Ross, T Vidaurre, HL Gomez, GN Hortobagyi, and L Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology,24(26):42364244*, 2006.

[36] Euler G. Horta. Previsores para a eficiência da quimioterapia neoadjuvante no câncer de mama. Dissertação de mestrado, Universidade Federal de Minas Gerais, 2008.

[37] Kaizhu Huang, Haiqin Yang, I. King, and M.R. Lyu. Maximin margin machine: Learning large margin classifiers locally and globally. *Neural Networks, IEEE Transactions on*, 19(2):260–272, Feb. 2008.

[38] P. Korhonen. Multiple objective programming support., 1998.

[39] A. Lima. Máquinas de vetores suporte na classificação de impressões digitais. Dissertação de mestrado, Universidde Federal do Ceará, 2002.

[40] Ricardo Hiroshi Caldeira ; BRAGA A. P. MEDEIROS, T. H. ; TAKAHASHI. A new decision strategy in multi-objective training of artificial neural networks. In *European Symposium on Neural Networks, Brugges*, pages 555–560, Bruxelas, 2007. Proceedings of the European Symposium on Neural Networks, D-Side Publications.

[41] T. Medeiros. *Otimização Multiobjetivo e Aprendizado de Máquina.* Exame de qualificação para doutorado, Universidade Federal de Minas Gerais, 2006.

[42] J.M. Mendel and R.W. McLaren. *Adaptive, Learning, and Pattern Recognition Systems; Theory and Applications.* Academic Press, New York, 1970.

[43] D. Muchnik. Support vector machines for classification. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 2000.

[44] R. Natowicz, A. P. Braga, R. Incitti, E. G. Horta, R. Rouzier, T. S. Rodrigues, and M. A. Costa.

[45] Rene Natowicz, Roberto Incitti, Euler Guimaraes Horta, Benoit Charles, Philippe Guinot, Kai Yan, Charles Coutant, Fabrice Andre, Lajos Pusztai, and Roman Rouzier. Prediction of the outcome of preoperative chemotherapy in breast cancer by dna probes that convey information on both complete and non complete responses. *BMC Bioinformatics, 9:149*, march 2008.

[46] B. Pearlmutter. *Gradient descent: second order momentum and saturation error.* In J.E. Moody, S. Hanson, and R. Lippmann, editors, Advances in Neural information Processing Systems 2, pages 887-894. Morgan Kaufmann, 1992.

[47] M. Riedmiller. Advanced supervised learning in multi-layer perceptrons - from backpropagation to adaptive learning algorithms, 1994.

[48] M. Riedmiller and H. Braun. Rprop- a fast adaptive learning algorithm, 1992.

[49] M. Riedmiller and H. Braun. Rprop – description and implementation details, 1994.

[50] D.E.; Rumelhart and J.L. McClelland. *Parallel Distributed Processing, vol1: Foundations.* The MIT Press, 1986.

[51] Y. Sawaragi, H. Nakayama, and T. Tanino. *Theory of multiobjective optimization.* Academic Press, 1985.

[52] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:322–330, 1998.

[53] F.-M. Schleif, B. Hammer, and T. Villmann. Margin-based active learning for lvq networks. *Neurocomputing*, 70(7-9):1215 – 1224, 2007. Advances in Computational Intelligence and Learning - 14th European Symposium on Artificial Neural Networks 2006, 14th European Symposium on Artificial Neural Networks 2006.

[54] R. Semolini. Support vector machines, inferência transdutiva e o problema de classificação. Dissertação de mestrado, Universidade Estadual de Campinas, 2002.

[55] John Shawe-taylor and Nello Cristianini. Smola, bartlett, scholkopf, and schuurmans: Advances in large margin classifiers, introduction to large margin classifiers.

[56] A. Sideris and S.E. Castella. A proximity algorithm for support vector machine classification. *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference on*, pages 2433–2438, Dec. 2005.

[57] R.M. Simon, E.L. Korn, L.M. McShane, M.D. Radmacher, G.W. Wright, and Y. Zhao. Design and analysis of dna microarray investigations. Springer, 2004.

[58] A.J. Smola. *Learning with Kernels.* The MIT Press, 2002.

[59] Songbo Tan. Large margin dragpushing strategy for centroid text categorization. *Expert Syst. Appl.*, 33(1):215–220, 2007.

[60] S. Tong. Support vector machine active learning for image retrieval.

[61] S. Tong. Support vector machine active learning for music retrieval. *Advances in Neural Information Processing Systems*, pages 547–553, 2000.

[62] Lorenzo Torresani and Kuang C. Lee. Large margin component analysis. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1385–1392. MIT Press, Cambridge, MA, 2007.

[63] B.; Guyon I. Vapnik, V.N.; Boser. A training algorithm for optimal matgin classifiers. *Fifth Annual Workshop on Computational Learning Theory, pp.1-152. San mateo, CA.*, 1992.

[64] C. Vapnik, V.N.; Cortes. Support vector networks. *Machine learning, vol.20, pp273-297.*, 1995.

[65] Vladimir N. Vapnik.

[66] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag: New York, 1995.

[67] V.N. Vapnik. *Statistical learning theory*. Wiley: New York, 1998.

[68] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, 18:1473–1480, 2006.

[69] Daniel S. Yeung, Defeng Wang, Wing W. Ng, Eric C. Tsang, and Xizhao Wang. Structured large margin machines: sensitive to data distributions. *Mach. Learn.*, 68(2):171–200, 2007.

[70] Dong Yu and Li Deng. Large-margin discriminative training of hidden markov models for speech recognition. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 429–438, Washington, DC, USA, 2007. IEEE Computer Society.