Previsores para a Eficiência da Quimioterapia Neoadjuvante no Câncer de Mama

Euler Guimarães Horta

Previsores para a Eficiência da Quimioterapia Neoadjuvante no Câncer de Mama

Euler Guimarães Horta

Orientador: *Prof Dr Antônio de Pádua Braga*
Co-orientador: *Prof Dr René Natowicz*

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
Novembro/2008

Agradecimentos

Agradeço primeiramente a Deus que me deu inúmeras oportunidades em minha vida, aos meus pais e minha irmã que sempre me deram carinho e apoio em todos os momentos que precisei e que sempre me incentivaram a estudar e vencer desafios. A minha namorada pelo amor, paciência e compreensão. Ao Braga por esses vários anos de orientação e amizade. Ao Thiago, Marcelo, Cristiano e Francisco por todos os conselhos e ajuda. A todos os colegas do LITC pelos bate-papos e convivência. Ao René, Roberto e Roman que me deram a oportunidade de trabalhar nos primórdios desse projeto na França. A Benoît e Philippe que tanta força me deram no tempo em que estive em Paris. Ao curso de engenharia elétrica, a CAPES e a ESIEE que me deram a oportunidade de participar do Brafitec. Ao CNPq pelo financiamento tanto na iniciação científica quanto no mestrado.

Resumo

O câncer de mama é atualmente o segundo tipo de câncer mais freqüente no mundo, ficando atrás apenas do câncer de pulmão, e é o mais freqüente entre as mulheres. O tratamento consiste em quimioterapia pré-operatória, cirurgia e quimioterapia pós-operatória sendo muito doloroso e com grandes efeitos psicológicos na mulher. No caso de câncer operável, a quimioterapia pré-operatória tem como objetivo reduzir o tamanho do tumor ou mesmo eliminá-lo, porém esse tratamento funciona em poucos casos, sendo que a grande maioria dos pacientes não tem grande redução do tumor. Desta forma a grande maioria dos pacientes que se submetem a este tipo de tratamento passam por esse sofrimento de forma desnecessária, uma vez que o tumor poderia ser retirado sem o uso dessa quimioterapia. Desta forma faz-se necessário saber se a paciente será sensível à quimioterapia ou não, para que o médico possa escolher o melhor tratamento. Essa previsão de sensibilidade ao tratamento feita através de características clínicas geralmente não dá bons resultados. Com o surgimento nos últimos anos da possibilidade de medir a expressão dos genes surge a esperança de utilizar essa informação genômica para tentar descobrir se características genéticas das pacientes podem estar relacionadas ao funcionamento ou não do tratamento.

Neste trabalho é proposta uma nova maneira de selecionar os genes que são mais informativos quanto à sensibilidade ou não da paciente à quimioterapia pré-operatória e são propostos classificadores que tentarão prever se os pacientes terão resposta patológica completa ao tratamento ou não. A metodologia aqui utilizada foi comparada com dois importantes trabalhos da literatura e mostrou ter melhores resultados, tanto para a previsão da eficácia da quimioterapia pré-operatória, quanto para o número de genes utilizados na classificação, que foi quase três vezes menor.

As principais características da abordagem utilizada para a seleção dos genes são o uso da técnica *Volcano Plot* para a seleção de um primeiro conjunto significativo de genes, o uso da AAC (area above the ROC) através do

classificador *naïve Bayes* que conseguiu refinar a seleção de genes e por fim o uso do método LASSO para treinamento multiobjetivo de redes neurais para um último refinamento da seleção de sondas. Utilizando essas três técnicas foi possível selecionar dois conjuntos de genes, um com 18 sondas e outro com 11 sondas (contido no primeiro), sendo que o primeiro fornece uma classificação melhor que a proposta na literatura e a segunda consegue aproximadamente a mesma performance que a literatura, porém com quase um terço dos dados utilizados.

Abstract

Breast cancer is the second most common kind of cancer in the world, behind only of lung cancer, and it is the most common among women. Treatment consists in pre-operative chemotherapy, surgery and post-operative chemotherapy and it is very painful and with great psychological effects in women. In the case of operable cancer pre-operative chemotherapy aims to reduce the size of the tumor or even eliminate it, but this treatment works in a few cases, where the vast majority of patients do not have large reduction of the tumor and they suffer unnecessary, since the tumor could be removed without the use of chemotherapy. Thus it is necessary to know if the patient is sensitive to chemotherapy or not, and the doctor can choose the best treatment. This prediction of sensitivity to the treatment made by clinical features usually does not give good results. With the possibility in recent years to measure the expression of genes it is desired to use that information to discover if some genetic feature is related or not to the functioning of the treatment.

In this work is proposed a new way to select genes that are more informative about the sensitivity of the patient or not to pre-operative chemotherapy and it is proposed classifiers that try to predict whether the patients will have complete pathological response to treatment or not.

Sumário

Lista de Abreviaturas

Abreviatura	**Significado**
PCR	Resposta Patológica Completa (*Pathologic Complete Response*)
RD	*Residual Disease*
KNN	*k-Nearest Neighbor Algorithm*
SVM	Máquinas de Vetores de Suporte (*Support Vector Machine*)
DLDA	*Diagonal Linear Discriminant Analysis*
AUC	Área Abaixo da Curva ROC (*Area Under de ROC Curve*)
AAC	Área Acima da Curva ROC (*Area Above de ROC Curve*)
MOBJ	Algoritmo de Treinamento Multiobjetivo de Redes Neurais
LASSO	*Least Absolute Shrinkage and Selection Operator*
MLP	Perceptron multi-camadas (*Multi-Layer Perceptron*)
DNA	Ácido Desoxirribonucleico (*deoxyribonucleic acid*)
cDNA	DNA Complementar
RNA	Ácido Ribonucleico (*ribonucleic acid*)
mRNA	RNA Mensageiro
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
FP	Falso Positivo
FN	Falso Negativo

Lista de Símbolos

Símbolo	**Significado**
$\mu(S)$	Média dos níveis de expressão de uma sonda
$\sigma(S)$	Desvio padrão dos níveis de expressão de uma sonda
$K(x, x)$	Kernel
w	vetor de pesos de uma rede neural
Q_k	Elipsoide
M	Matriz de gradientes da restrições
P	Matriz de projeção

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

O câncer de mama é o tipo de tumor mais freqüente na mulher, atingindo-a com maior freqüência após os 40 anos de idade, apesar de, nos últimos anos ter se observado um fenômeno em nível mundial, ainda inexplicado, que é o aumento sensível de sua incidência em faixas etárias mais jovens [17]. As estimativas do número de casos novos de neoplasias na população feminina para o Brasil em 2008, apontam o câncer de mama em primeiro lugar com 49.400 novos casos [13]. Trata-se de uma doença heterogênea e complexa, como observado pelas múltiplas formas de apresentação clínica e morfológica, bem como pelas diferenças na pré e pós-menopausa, pelos diferentes graus de agressividade tumoral e pelo potencial metástico.

Uma das maiores dificuldades no tratamento está no fato de que, quando o tumor atinge um limiar de detecção clínica, em torno de 1 cm, este apresenta uma massa celular de aproximadamente 109 células e pesa cerca de 1g, tendo duplicado 30 vezes em média, com um tempo médio de duplicação que varia de 30 a 200 dias. Desse modo, um tumor considerado clinicamente precoce já existiu em fase pré-clínica por um período de 2 a 17 anos, tendo evoluído 3/4 de sua vida biológica antes de causar a morte da paciente [17] [4].

A observação das taxas de cura e sobrevida aos 10 anos tem demonstrado que o potencial metástico, muitas vezes, já se manifestou antes do diagnóstico clínico. Embora células cancerígenas possam ser liberadas do tumor antes do diagnóstico, variações na capacidade de o tumor crescer em outros órgãos e a resposta do hospedeiro ao tumor podem inibir a disseminação da doença.

O tratamento quimioterápico pré-operatório é uma das abordagens mais disseminadas atualmente. A correta previsão da sensibilidade do paciente para tal quimioterapia pré-operatória é uma questão vital no tratamento do

câncer de mama: pacientes que apresentam uma resposta patológica completa - *PCR* a tal tratamento devem ser previstos com a maior acurácia possível a fim de que se investiguem outras possibilidades para os pacientes que não apresentam uma resposta patológica satisfatória.

Existem na literatura basicamente duas formas de previsão de PCR, uma através de informações clínicas dos pacientes e outra através de informações de níveis de expressão dos genes dos pacientes. Muitos trabalhos tem sido desenvolvidos na área de previsão de PCR através dos níveis de expressão dos genes [23, 19], sendo que o que obteve melhores resultados até o presente momento foi o trabalho apresentado por Natowicz et al. [27]. Nesse trabalho Natowicz et al. propuseram uma nova metodologia para seleção de sondas de *microarray* bem como um novo modelo de classificador. Natowicz et al. utilizaram o conjunto de teste para selecionar o melhor classificador, o que pode ter feito que seu modelo tenha ficado super-ajustado (*overfitting*) ao conjunto de teste, podendo ter resultado em um modelo com baixa capacidade de generalização. Uma forma que poderia ter sido utilizada para garantir uma boa generalização seria utilizar uma pequena parte do conjunto de treinamento como conjunto de validação para selecionar o modelo e, após isto, utilizar o conjunto de teste para verificar a performance do classificador proposto. Tal abordagem não foi utilizada devido a escassez dos dados. Nesta dissertação será proposta uma forma para resolver o problema utilizando a técnica de validação cruzada.

Uma das dificuldades para realizar a previsão de PCR através de dados de níveis de expressão de genes é a grande dimensão das bases de dados, pois para cada paciente são geradas as expressões de mais de 22000 sondas, sendo que uma ou mais sondas são utilizadas para "apontar um gene´´, o que torna necessária a realização de uma redução de dimensão através da seleção das sondas mais informativas. Outra dificuldade é o custo para a obtenção dos dados, uma vez que para gerar o *microarray* de apenas um paciente é gasto em torno de mil dólares. Dessa forma o problema tem um grande espaço de entrada e poucos dados para a análise. A questão da previsão de PCR impõe ainda uma terceira dificuldade, uma vez que as classes (PCR e NoPCR) são totalmente desbalanceadas, já que em aproximadamente 70% dos casos os pacientes são NoPCR (não há resposta patológica completa).

Uma das vantagens em reduzir o espaço de entrada é, além de facilitar a análise, possibilitar que no futuro *microarrays* menores possam ser utilizados como um exame para a verificação da sensibilidade à quimioterapia neoadjuvante, o que vai baratear o custo de produção do *array*, utilizando, por exemplo, uma centena de sondas no lugar de milhares de sondas.

Dessa forma essa dissertação tem três objetivos: selecionar o menor con-

junto possível de sondas significativas; ajustar um classificador com boa capacidade de generalização utilizando os dados de treinamento disponíveis, testando a generalização com o conjunto de teste; tentar contornar o problema de desbalanceamento das classes.

1.1 Contribuições

Neste trabalho é utilizada uma técnica consagrada para a seleção de sondas significativas, denominada *Volcano Plot*. Essa ferramenta é resultado da união de duas técnicas famosas na área de análise de níveis de expressão, sendo a *fold change* e o teste t. Com essa ferramenta foi possível selecionar 81 sondas significativas o que já foi um grande ganho se comparado com o tamanho do espaço de entrada (22283 sondas). Dentre as 81 sondas selecionadas existia redundância, uma vez que mais de uma sonda identificava o mesmo gene. Assim as sondas redundantes foram filtradas resultando em um conjunto de 62 sondas.

Para reduzir ainda mais o espaço de entrada foi proposta nessa dissertação uma técnica que utilizará o classificador *naïve Bayes*, validação cruzada e a avaliação de performance através da AAC (area above the ROC). Com essa técnica foi possível obter um espaço de entrada de apenas 18 sondas que tem performance de classificação superior às técnicas propostas na literatura.

Na dissertação vários classificadores foram testados. O classificador baseado no treinamento multiobjetivo de redes neurais artificiais pelo método LASSO forneceu como efeito "colateral" uma maior redução do espaço de entrada, dando resultados semelhantes aos da literatura, porém utilizando apenas 11 sondas.

Por fim essa dissertação avaliou a performance desses dois conjuntos de sondas selecionados do ponto de vista dos classificadores: voto majoritário; DLDA; naïve Bayes; SVM; comitê de perceptrons; redes neurais MLP; redes neurais MOBJ; redes neurais LASSO.

1.2 Organização do texto

A dissertação está dividida em 8 capítulos da seguinte forma:

- No capítulo 2 são apresentados os conhecimentos necessários para o entendimento da dissertação. O problema do câncer e mais especificamente câncer de mama é apresentado. Conceitos básicos de genética também são abordados. O funcionamento geral da tecnologia de *microarrays* e o funcionamento dos *microarrays Affymetrix* também são apresentados.

- No capítulo 3 é apresentado o problema principal tratado na dissertação, a previsão de resposta patológica completa para a quimioterapia neoadjuvante, mostrando as dificuldades inerentes ao problema e os trabalhos relevantes dessa área que foram utilizados nesse texto.

- No capítulo 4 é apresentada a base de dados utilizada na dissertação e são apresentadas as duas principais técnicas para a seleção de sondas que já foram aplicadas ao problema de previsão de PCR na literatura e é apresentada ainda a técnica de seleção de sondas proposta nesta dissertação.

- No capítulo 5 é feita uma rápida introdução aos classificadores utilizados neste trabalho, a saber: voto majoritário; DLDA; naïve Bayes; SVM; comitê de perceptrons; redes neurais MLP; redes neurais com treinamento MOBJ; redes neurais com treinamento LASSO.

- No capítulo 6 são apresentados os resultados da performance de cada classificador para cada conjunto de sondas selecionadas. Ao final do capítulo é feita a avaliação dos resultados.

- No capítulo 7 é feita uma discussão sobre os conjuntos de sondas selecionadas obtidos na dissertação em relação aos conjuntos de sondas da literatura. São discutidas ainda algumas características biológicas dos genes selecionados.

- No capítulo 8 são feitas as conclusões finais da dissertação e são apresentadas as possibilidades de trabalhos futuros.

1.3 Conclusões do capítulo

Neste capítulo foi dada uma visão geral do problema tratado nessa dissertação, mostrando as dificuldades e os desafios encontrados. Foram apresentados os três objetivos principais da dissertação que são: selecionar o menor conjunto possível de sondas significativas; ajustar um classificador com boa capacidade de generalização utilizando os dados de treinamento disponíveis, testando a generalização com o conjunto de teste; tentar contornar o problema de desbalanceamento das classes. Por fim, foi apresentada uma visão geral de cada capítulo que compõem esse trabalho.

No capítulo seguinte serão apresentados os conceitos básicos necessários para o entendimento do trabalho apresentado nessa dissertação. Será apresentada a doença (câncer de mama), noções de genética humana e o funcionamento e tipos de *microarray*.

CAPÍTULO 2

Câncer de Mama, Genética Humana e *Microarray*: Conceitos Básicos

O cancêr de mama é o tumor que mais atinge as mulheres no mundo, tornando-se um problema de saúde pública. Seu tratamento é realizado geralmente em três fases: quimioterapia neoadjuvante; cirurgia e quimioterapia adjuvante. Um dos problemas dessa seqüência de tratamentos é que a quimioterapia neoadjuvante funciona apenas em aproximadamente 30% dos pacientes, sendo que em torno de 70% dessas pacientes sofrem os efeitos negativos da quimioterapia inutilmente.

Um grande desejo dos médicos é conseguir prever se a quimioterapia neoadjuvante vai dar bons resultados ou não a fim de evitar o sofrimento desnecessário das pacientes. Para tanto, muitos trabalhos foram desenvolvidos com esse objetivo, sendo que em sua grande maioria foram utilizadas informações clínicas para realizar essa previsão. Essa abordagem, porém, ainda não tem dado bons resultados. Com o surgimento da tecnologia de *microarrays* uma nova esperança surgiu, já que com essa tecnologia seria possível verificar se o paciente tem uma tendência genética para uma boa resposta a esse tratamento.

Esse capítulo apresentará algumas considerações e características gerais da doença, bem como os conhecimentos básicos de genética humana necessários para o entendimento do funcionamento dos *microarrays*.

2.1 Câncer de mama

O corpo humano é constituído de pequenas células que crescem e morrem de forma controlada. Quando essas células crescem e se dividem de forma descontrolada e anormal surgem os tumores. Se o tumor não invade os tecidos vizinhos e nem outras partes do corpo ele é chamado benigno ou não-canceroso e geralmente não causa risco de morte. Caso o tumor invada as células vizinhas e as destruam ou caso ele se espalhe pelo corpo ele passa a ser chamado de tumor maligno ou câncer. Esse tipo de tumor coloca em risco a vida do paciente [28].

As células contêm material genético que é responsável pelo controle do crescimento celular. Alterações neste material podem acarretar na perda do controle do crescimento celular, o que pode provocar o câncer.

O câncer de mama é um tumor maligno que começa nas células das mamas. Este tipo de câncer ocorre mais comumente nas mulheres, mas pode afetar também os homens (1 homem para cada 100 mulheres) [36].

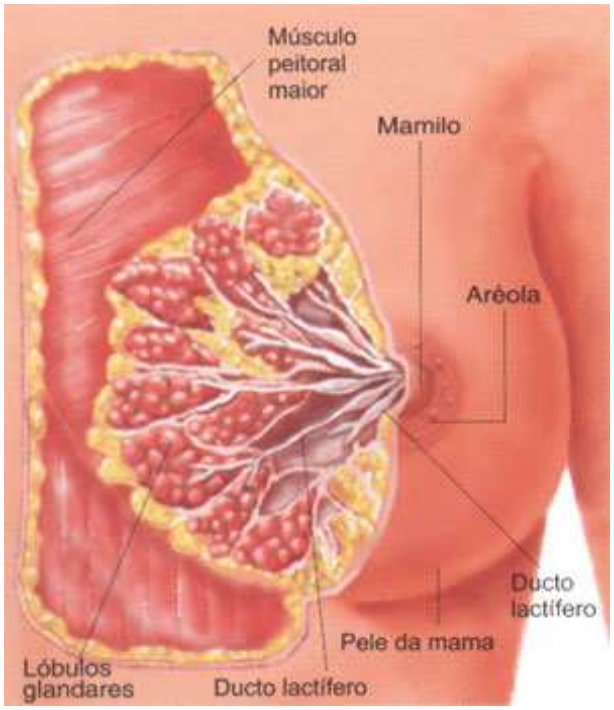


Figura 2.1: Anatomia da mama [9]

A Figura 2.1 mostra a anatomia da mama. Nela são apresentadas as estruturas que são responsáveis pela produção do leite (lóbulos glandares) e pela condução do mesmo até o mamilo (ducto lactífero). A forma mais comum do câncer de mama é a que atinge os ductos lactíferos, chamado Carcinoma Ductal. Ele pode ser *in situ*, quando não passa das primeiras camadas de célula destes ductos, ou invasor, quando invade os tecidos em volta. O tipo de câncer que começa nos lóbulos da mama é chamado de Carcinoma Lobular e é menos

comum que o primeiro. Este tipo de câncer 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 [14].

2.1.1 Sintomas

Em geral o câncer de mama não dói. O principal sintoma é um nódulo (ou caroço) que surge na mama e é sentido pela paciente quando ela toca a região das mamas. Caso esse nódulo apareça a paciente deve procurar rapidamente um médico para que ele faça o exame de toque. O exame consiste em palpar as mamas, as axilas, a região do pescoço e clavículas. Caso ele confirme a presença do nódulo ele encaminha a paciente para uma mamografia (um exame de raio X das mamas) [14].

A mulher pode perceber ainda assimetrias ou deformidades nas mamas, perceber retração na pele. Pode aparecer também um líqüido sanguinolento saindo pelo mamilo. Em casos mais avançados da doença podem surgir uma ferida (ulceração) na pele[14].

No caso de carcinoma inflamatório a mama pode aumentar rapidamente de volume, ficando quente e vermelha[14].

2.1.2 Fatores de risco

Os principais fatores de risco que podem aumentar a chance de uma mulher desenvolver o câncer de mama são [14]:

- Idade superior a 50 anos;
- Exposição excessiva a hormônios, como o uso prolongado de anticoncepcionais orais (pílulas) ou a terapia de reposição hormonal baseada em hormônios femininos como o estrogênio e a progesterona;
- Exposição a radiação na região do tórax;
- Ingestão excessiva de bebidas alcoólicas;
- Obesidade;
- Início da menstruação antes dos 11 anos ou menopausa tardia;
- Ter tido um câncer anterior.

2.1.3 Tratamento

O tratamento do câncer de mama é constituído geralmente de 3 etapas: quimioterapia neoadjuvante, cirurgia e quimioterapia adjuvante. As três etapas são descritas a seguir.

Quimioterapia Neoadjuvante

A quimioterapia neoadjuvante é o tratamento feito com drogas quimioterápicas antes da cirurgia visando diminuir o tumor (redução parcial) de forma a viabilizar que parte da mama que esteja livre da doença não seja retirada. Ela pode ainda extinguir o tumor (resposta patológica completa). É utilizada geralmente em pacientes com câncer localmente avançado.

Cirurgia

Podem ser realizados dois tipos de cirurgia, dependendo do tumor da paciente. O primeiro tipo é a cirurgia conservadora indicada para pacientes que tem partes da mama livres da doença e o segundo tipo é a mastectomia que é a retirada de toda a mama da paciente sendo a cirurgia mais radical e que tem fortes impactos psicológicos nas pacientes.

Quimioterapia Adjuvante

A quimioterapia adjuvante é o tratamento feito com drogas quimioterápicas após a cirurgia visando que o tumor não volte e para evitar que a paciente sofra metástase.

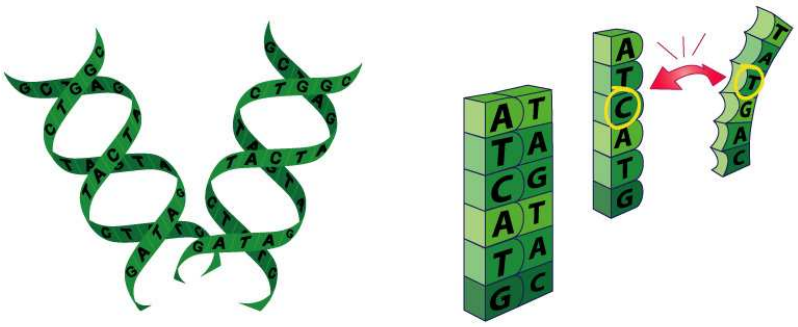
2.2 Conceitos de genética humana

O DNA é uma cadeia de moléculas em forma de dupla hélice. Ele contém toda a informação necessária para todos os processos celulares. O DNA funciona como um "*script*" para criar o ácido ribonucléico (RNA) e as proteínas, sendo que o RNA coordena a produção das proteínas [1].

A cadeia de DNA é constituída por apenas quatro tipos de moléculas: adenina (A), guanina (G), citosina (C) e timina (T). Elas são chamadas de *bases*. Essas bases se organizam em pares, sendo que a adenina se conecta apenas com a timina e a guanina se conecta apenas com a citosina. Desta forma uma fita de DNA se conectará a outra para formar a hélice somente se uma for completar a outra. Por exemplo a fita ATCATG só se conectará à fita TAGTAC, já que todas as bases de uma são complementares às bases da outra. Se pelo menos uma das bases não for complementar na segunda fita as duas vão se

repelir. A diferença básica entre o DNA e o RNA é que este possui a uracila (U) no lugar da timina (T). O RNA funciona como um mensageiro, carrengando em apenas uma fita simples de bases a informação necessária para produzir uma proteína, ou seja, o RNA funciona como um "*template*" para a produção de uma proteína. A Figura 2.2 mostra um exemplo de hélice e de pares de bases que compõem o DNA.

Um gene é definido com sendo uma parte do DNA que contém a informação para produzir uma certa proteína. Para se criar uma proteína primeiramente as duas cadeias simples da dupla hélice do DNA se afastam. Moléculas de RNA são atraídas para a cadeia e nela se conectam formando uma imagem espelhada da cadeia, porém com uracila no lugar da timina. Esse processo é denominado transcrição. Após concluir a cópia, o RNA se desconecta da cadeia e se dirige ao citoplasma onde o ribossomo sintetiza a proteína no processo denominado translação. As cadeias do DNA se reconectam e voltam para a forma original de dupla hélice. A figura 2.3 mostra resumidamente este processo.



(a) Exemplo de dulpa hélice formadora do DNA.

(b) Pares de bases e repulsão de bases não complementares

Figura 2.2: Dupla hélice e pares de bases. Fonte: Site da Affymetrix [1]

2.3 Microarray

Nos últimos 15 anos a tecnologia de *microarray* tem evoluído rapidamente. Com o avanço da tecnologia o custo para produzir os *arrays* tem diminuído e a capacidade de armazenamento e análise de dados tem aumentado bastante, tornando viáveis aplicações práticas como a análise dos genes que estão correlacionados com doenças como o câncer [37].

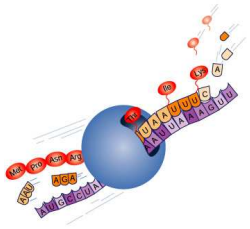
A revolução causada por essa tecnologia começou em 1995 com o trabalho de Mark Schena et al [33] em que os autores apresentaram uma técnica baseada em DNA complementar (cDNA) para medir a expressão de genes. Para

(a) Gene

(b) A dupla hélice se separa

(c) O RNA começa a ser gerado na transcrição

(d) O RNA se desprende no final da transcrição



(e) O RNA é utilizado como *template* para a criação de uma proteína no ribossomo

Figura 2.3: Processos de transcrição e translação. Fonte: Site da Affymetrix [1]

tanto eles colocaram cDNA de vários genes do organismo *Arabidopsis thaliana* em uma lâmina de microscópio, utilizando para isso um robô. Posteriormente foram preparadas sondas ou *probes* fluorescentes de mRNA (RNA mensageiro) para cada gene que eles desejavam analisar. Os autores previram ainda que os *arrays* comportariam facilmente mais de 20000 genes, o que é realidade hoje nos *microarrays* affymetrix. A tecnologia *Affymetrix* foi apresentada na *Nature Biotechnology* por D. J. Lockhart et al. [22] em 1996.

Essa tecnologia é capaz de analisar vários genes em paralelo, o que possibilita um entendimento maior dos mecanismos responsáveis por características ou mesmo por doenças. Desta forma seu uso é promissor no estudo do câncer de mama ou em outros tipos de câncer.

Um *microarray* de DNA consiste de uma superfície sólida, geralmente um *slide* de microscópio, no qual são depositadas moléculas de cDNA. O objetivo da técnica é detectar a presença e a quantidade de ácido nucléico rotulado em amostras biológicas, como tecidos, sangue, etc, medindo assim o nível de expressão de determinados genes. A grande vantagem da técnica é que pode-se medir a expressão de milhares de genes simultaneamente [37].

2.4 Funcionamento geral dos Microarrays

O uso dos *microarrays* é realizado em seis passos [37]:

- Escolha dos tecidos a serem estudados
- Extração de RNA
- Preparação da amostra e rotulamento dos genes procurados
- Hibridização
- Lavagem
- Aquisição da imagem

A figura 2.4 mostra esquematicamente o uso de microarrays. Cada microarray contém partes dos genes que se deseja medir a expressão no experimento, desta forma quem for realizar o experimento deve saber exatamente que genes serão analisados.

O primeiro passo é escolher quais tecidos serão comparados. A tecnologia de microarray consiste em uma forma de comparar a expressão gênica entre dois tecidos, sendo um tecido normal e outro a ser analisado, que neste caso é o tecido doente. O segundo passo consiste em extrair o RNA das amostras dos dois tecidos e aplicar um marcador (rótulo) fluorescente. O marcador Cy3 fornece a cor verde à amostra e o marcador Cy5 fornece a cor vermelha. Feito isso as amostras são colocadas no *microarray* e sofrem hibridização. O processo de hibridização consiste na combinação do RNA com a fita do cDNA presente no array. Como visto anteriormente o RNA só se combinará com o cDNA de bases complementares. Quando essa combinação ocorre, forma-se uma hélice e o marcador fluorescente permanece conectado na nova hélice. O processo de hibridização é mostrado na figura 2.5.

Após este passo é feita a lavagem da lâmina, a fim de retirar as amostras que não sofreram hibridização, ou seja, que não se combinaram com o DNA presente no array. O array passa então por um *scanner* que emitirá um feixe de laser vermelho e um feixe de laser verde, que ativarão a fluorescência dos marcadores. O *scanner* fará a aquisição da imagem que será tratada posteriormente, com técnicas de *clutering* e análise de imagem sendo os resultados normalizados ao final para a criação de um vetor numérico que representará a expressão dos genes.

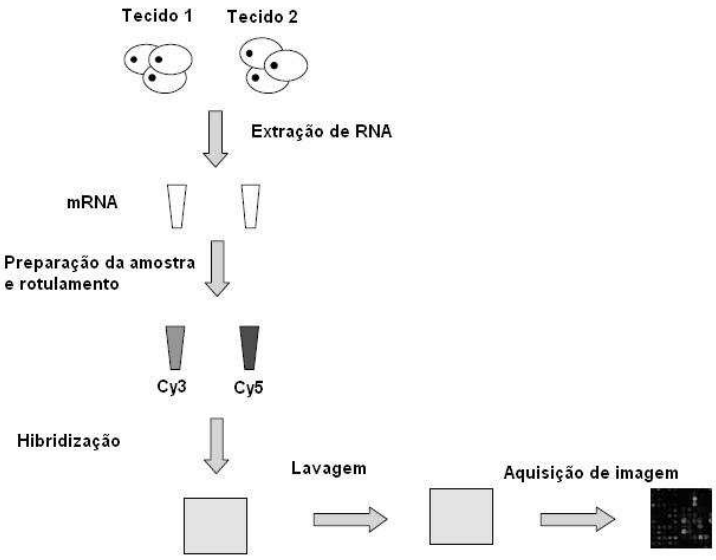


Figura 2.4: Os 6 passos do uso de microarrays. Fonte: Microarray Bioinformatics [37]

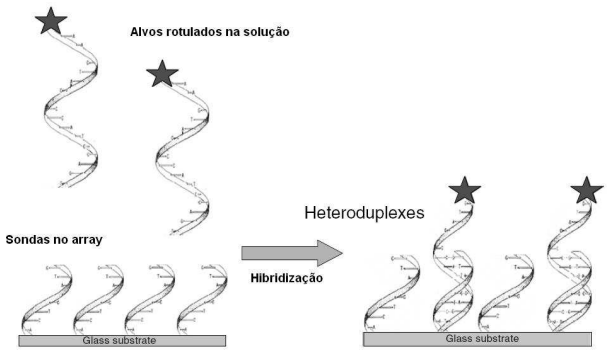


Figura 2.5: Hibridização. Fonte: Microarray Bioinformatics [37]

2.5 Microarray Affymetrix

O tipo de *microarray* utilizado nesta dissertação foi o *Affymetrix*. Algumas particularidades deste tipo de array serão tratadas nessa seção.

Assim como todos os tipos de *microarray* o *Affymetrix* também se baseia na hibridização para medir a expressão dos genes.

Esse técnica pode medir a expressão de todos os genes humano conhecidos, ou de qualquer outro organismo que se deseje.

O primeiro passo é colocar um trecho de DNA, aqui chamada de *probe* ou sonda, na superfície do *chip* (ou *slide*) de vidro. Este filamento tem apenas 25 bases de comprimento que representa uma pequena seção de um gene muito maior. Os cientistas comparam a sonda de seqüência de 25 bases com o resto do genoma humano para ter certeza que este trecho do gene não se repete em nenhum outro gene. Desta forma, quando uma molécula de RNA encontra a sonda, os pesquisadores vão saber que o gene foi expresso, porque

aquela seqüência ocorre somente para o gene especificado. Assim este trecho de DNA mede a expressão do gene completo [1]. Para aumentar a exatidão dos resultados gerados pelos *microarrays* são utilizados 11 seções diferentes do mesmo gene para sua identificação, o que implica em 11 sondas para cada gene medido. Diferentemente dos outros tipos de *microarray*, o affymetrix é mono-cor, uma vez que não necessita de um técido de referência, já que as sondas já contém os genes do genoma humano. Desta forma o resultado da análise é o nível de expressão de cada gene do tecido estudado e não a relação entre níveis de expressão entre tecido alvo e tecido de referência.

2.6 Conclusões do capítulo

Neste capítulo o câncer de mama foi caracterizado e foram apresentadas as suas principais características. Conceitos de genética humana foram revistos para darem suporte ao entendimento do funcionamento dos *microarrays*. Por fim foram apresentados o funcionamento geral dos *microarrays* e as características principais da tecnologia *Affymetrix*.

No capítulo seguinte será apresentado o problema de previsão de resposta patológica completa para a quimioterapia neoadjuvante e os principais trabalhos dessa área.

CAPÍTULO 3

Estado da Arte

A previsão de resposta patológica completa (PCR) para a quimioterapia neoadjuvante é um tema de grande interesse da comunidade científica. Esse interesse se dá ao fato de esse tratamento funcionar em apenas aproximadamente 30% dos casos, de modo que o paciente pode sofrer os efeitos colaterais desnecessarimante. Com a possibilidade de medir o nível de expressão de genes, muitos trabalhos tem tentado realizar essa previsão baseada nesse tipo de dados.

As principais dificuldades que os pesquisadores tem enfrentado nessa área são o grande espaço de busca (mais de 20000 sondas por paciente), o pequeno número de pacientes nas bases de dados, devido ao alto custo para a obtenção dos dados e a dificuldade de obter uma boa generalização dos modelos, grande parte devido a essa escassez de dados.

3.1 Caracterização do problema

Como visto no capítulo anterior, a quimioterapia neoadjuvante é o tratamento utilizado antes da cirurgia, visando acabar com o tumor ou pelo menos reduzi-lo, para que a cirurgia possa ser conservadora. O problema é que ela não funciona na maioria das vezes, fazendo com que o paciente sofra inutilmente.

Ao longo do tempo observou-se que tentar prever a eficácia do tratamento somente utilizando dados clínicos não dava bons resultados. Isso motivou os pesquisadores a buscar na tecnologia de *microarrays* um novo caminho para essa previsão. Essa nova abordagem tem como vantagem possibilitar que o problema da ineficiência desse tipo de quimioterapia seja entendido

a fundo, em nível do perfil genético do paciente. Essa abordagem também trouxe consigo o problema da dimensão do espaço de busca, uma vez que, para cada paciente, mais de 20000 sondas são avaliadas no *array*. Dessa forma a estatística e a inteligência computacional mostram-se ferramentas promissoras para a analise desses dados. Outro problema inerente a essa tecnologia é o custo de produção do *array*, sendo que o exame pode chegar a 1000 dólares por paciente. Esse elevado custo traz consigo a dificuldade em obter dados para pesquisa.

Como a quimioterapia neoadjuvante funciona somente em torno de 30% dos casos o problema de previsão de resposta patológica completa é desbalanceado, o que dificulta ainda mais sua análise. Essa dificuldade ocorre também quando tenta-se treinar algum classificador, podendo o modelo ficar superajustado aos dados de treinamento.

Dessa forma devem ser tratados três problemas: redução do espaço de busca, selecionando somente as sondas mais significativas; evitar superajuste dos modelos de classificadores aos dados de treinamento; contornar o problema do desbalanceamento de classes.

A seguir serão apresentados os principais trabalhos nessa área, que vão ser de grande importância para o desenvolvimento dessa dissertação. São apresentados ainda alguns trabalhos do grupo de pesquisa do qual faz parte o autor desse texto.

3.2 Principais trabalhos

Em 2005 Olga Modlich et al. [23] realizaram uma seleção de genes que estão altamente relacionados com a presença ou não da doença, sendo os genes FHL1 e CLDN5, uma vez que eles são altamente expressos em tecidos normais e fracamente expressos em tecidos de tumores, sendo que são raramente detectados em tecidos de câncer de mama. Os autores também selecionaram dois conjuntos de genes para a previsão da eficiência da quimioterapia neoadjuvante utilizando algumas técnicas estatísticas. Eles dividiram os pacientes em PCR (*pathologic complete response*), NC (not change) e PR (partial reduction). Para a classificação foram utilizadas o KNN e PLS-DA (Partial least squares discriminant analysis). Como resultados os autores obtiveram especificidade média de mais de 74% para classe pCR, 62% para NC e 100% para PR. Eles explicaram ainda que PCR (*pathologic complete response*) não é sinônimo de cura, uma vez que mesmo que o tumor tenha desaparecido ainda existe a possibilidade de que celulas de metástase tenham se espalhado pelo corpo do paciente e dizem ainda que apesar disso a previsão de pCR pode ser um bom indicativo de sobrevivência após o tratamento.

K. Hess et al [19] apresentaram em setembro de 2006 o primeiro trabalho de desenvolvimento de previsores multigênicos para quimioterapia pre-operatória sobre os dados utilizados nesta dissertação. Os dados são compostos de *oligonecleotide microarrays* sendo 82 pacientes para treinamento e 51 para validação. O objetivo desse trabalho foi desenvolver um previsor de sensibilidade à quimioterapia com Paclitaxel e Fluorouracil-doxorubicin-cyclophsphamide.

Os dados de microarray foram normalizados utilizando o software dCHIP V1.3 (http://www.dchip.org). A normalização é descrita por E. Schadt et al. [32].

Os autores avaliaram 780 classificadores diferentes, dentre eles knn (k-nearest neighbor), SVM (support vector machine) e DLDA (Diagonal Linear Discriminant Analysis). Além de variar o tipo de classificador eles também variaram a quantidade de sondas utilizadas em cada classificador. O critério utilizado para a escolha do melhor modelo foi a menor área acima da curva ROC. Segundo este critério o melhor classificador foi o DLDA com 30 *probes* tendo acurácia de 0.76, sensibilidade de 0.92 e especificidade de 0.71 sobre o conjunto de validação.

Os autores comparam os resultados com o classificador gerado a partir de dados clínicos utilizando a técnica do DLDA. As características utilizadas foram ER (estrogênio receptor), idade e *nuclear grade*. Para este classificador a acurácia foi de 0.78, a sensibilidade 0.61 e a especificidade 0.84 sobre o conjunto de validação. Eles tentaram ainda unir os dois classificadores, mas sem conseguir grandes melhorias.

Natowicz et al. [27] apresentaram uma nova técnica para seleção de sondas baseada em intervalos de valores para os níveis de expressão de cada classe estudada, ou seja, calculam-se intervalos de expressão para as classes PCR e NoPCR. Ele define o intervalo como sendo a média $\pm$ 1 desvio padrão dos níveis de expressão de cada sonda em cada classe. Natowicz considerou que cada sonda seria um classificador independente e que a classificação consistiria em verificar se o nível de expressão de uma sonda está ou não dentro do intervalo PCR ou NoPCR, sendo considerada como indeterminada em caso do valor estar fora dos intervalos ou na interseção. A classificação é realizada tomando o voto majoritário das classificações de cada sonda selecionada. As sondas foram selecionadas pelo poder de classificação definido por V(s) = 0.5(p/P + n/N), onde p e n são as classificações corretas sobre o conjunto de treinamento e P e N são o número de padrões de treinamento da classe PCR e NoPCR, respectivamente. As sondas escolhidas foram aquelas com maior V(s) e que resultavam na melhor classificação do conjunto de teste. Este modelo é considerado como o melhor existente até o momento.

Este trabalho desenvolvido por Natowicz et al. [27] deu origem a uma série de trabalhos desenvolvidos em uma parceria entre o grupo de pesquisadores franceses e o grupo de pesquisadores brasileiros do qual faço parte, dentro do projeto CAPES-COFECUB. O primeiro trabalho desta série foi apresentado no ESANN 2008 (European Symposium on Artificial Neural Networks) [24] em que as equipes utilizaram o modelo de treinamento multiobjetivo de redes neurais (MOBJ) utilizando como espaço de entrada as 30 sondas selecionadas no trabalho anterior. Nesse trabalho foi observado que o modelo multiobjetivo melhora um pouco os resultados, mas a solução varia muito de acordo com o decisor utilizado.

O segundo trabalho da série foi apresentado por Braga et al. [7] no congresso ANNPR 2008 (Artificial Neural Networks in Pattern Recognition). Nesse trabalho os autores aplicaram o modelo *naïve Bayes* nas trinta sondas selecionadas a fim de investigar a hipótese de que o classificador de voto majoritário seria na verdade uma variação sem *priors* daquele classificador. Os autores testaram quatro variações para esse modelo, sendo uma com *priors* e dados discretizados (utilizando a classificação de cada sonda em PCR (1), NoPCR (-1) e indeterminado (0)), a seguinte sem *priors* e dados discretizados e as duas últimas com dados de nível de expressão brutos com e sem *priors*. Foi observado que o modelo proposto por Natowicz et al. [27] tende ao modelo *naïve Bayes* com dados discretizados e sem *priors*.

O terceiro trabalho da série foi apresentado no KES 2008 (Knowledge-Based and Intelligent Information & Engineering Systems) [26]. Nesse trabalho Natowicz et al. verificaram que com apenas 17 das 30 sondas era possível obter uma boa classificação dos pacientes utilizando para isso SVM com kernel linear. Esse trabalho foi escolhido para um artigo estendido em formato de capítulo de livro [25].

O último trabalho dessa série foi um capítulo publicado no livro "*Learning and Approximation: Theoretical Foundations and Applications*´´ por Costa et al. [11]. Nesse trabalho os autores utilizaram as trinta sondas selecionadas por Natowicz et al. [27] e normalizaram os níveis de expressão de forma que para cada sonda a média fosse zero e o desvio padrão um. Os dados de treinamento normalizados foram apresentados para os algoritmos MOBJ [39, 16] e LASSO [10]. Para os dois modelos as redes foram treinadas dez vezes e para cada execução a solução do Pareto foi escolhida utilizando um decisor de intersecção [24]. A solução escolhida entre as 10 execuções foi aquela que apareceu mais vezes. Foi observado que os resultados obtidos pelos dois métodos foram muito próximos, mas o método LASSO teve a vantagem de reduzir muito a complexidade da rede, selecionando os atributos de entrada mais importantes e reduzindo a camada escondida para apenas um neurônio.

3.3 Conclusões do capítulo

Nesse capítulo o problema de previsão de PCR foi caracterizado e foram apresentados os principais trabalhos dessa área. Esses trabalhos formam a base utilizada para o desenvolvimento desta dissertação. Foram apresentados ainda uma série de trabalhos realizados pelo grupo de pesquisa do qual o autor desta dissertação faz parte.

CAPÍTULO 4

Metodologia

A tecnologia de *microarrays* possibilita que uma grande quantidade de genes possa ser analisada simultaneamente. Isso fornece uma grande quantidade de informação que precisa ser analisada. Cada gene é representado no *array* por mais de uma sonda (ou *probe*), uma vez que deseja-se uma boa precisão do valor de expressão de cada gene. Em geral nem todas as sondas trazem informações pertinentes ao problema estudado. Desta forma torna-se necessário o uso de alguma técnica de seleção de dados que seja capaz de identificar quais sondas são mais importantes e estão mais relacionadas ao problema estudado.

A redução da dimensão do espaço de busca também é importante visando a otimizar o custo de produção dos *arrays*, uma vez que produzir um array com uma centena de sondas é muito mais barato que produzir um *array* com milhares de sondas. Além disso ao encontrar os genes mais significativos é possível fazer uma melhor análise biológica, sabendo, com mais clareza, que genes e proteínas estão relacionadas com o problema.

Neste capítulo serão apresentadas a base de dados e três técnicas para a seleção de sondas, sendo que duas delas já foram aplicadas na base de dados e a terceira técnica será utilizada nesses dados pela primeira vez.

4.1 Base de dados

A base de dados utilizada é constituída por 133 pacientes, sendo 82 para treinamento e 51 para teste. Os dados dos pacientes de treinamento foram coletados em Houston, EUA, e os dados de validação foram coletados parte em Villejuif, França e parte no Peru. Os pacientes foram tratados durante

24 semanas com a quimioterapia Paclitaxel e Fluorouracil - doxorubicin - cyclophsphamide. Algumas características clínicas foram coletadas antes do tratamento, como idade, raça, tamanho do tumor, tipo de tumor, etc. Ao final do tratamento foi verificado se restaram células cancerígenas na mama das pacientes. Se sim o paciente era classificado como RD (*residual disease*), caso contrário o paciente era classificado como pCR (*pathologic complete response*). Desta forma RD significa que o paciente precisará passar por cirurgia, já que o tumor persiste e pCR significa que o paciente não tem carcinoma residual com componente invasiva, o que implica em desaparecimento do tumor. É importante ressaltar que pCR não implica em cura, uma vez que células de metástase podem ter se espalhado pelo corpo do paciente [23]. Entretanto, a literatura mostra que pacientes pCR tem uma sobrevida maior e com mais qualidade do que pacientes RD (aqui tratado como NoPCR).

A base de dados está dividida da seguinte forma:

- Dados dos 82 pacientes coletados em Houston nos EUA:
  - 61 são No-PCR e 21 são PCR.
- Dados dos 51 pacientes coletados em Villejuif na França:
  - 38 são No-PCR e 13 são PCR.

4.2 Seleção de sondas

4.2.1 Seleção baseada no ranking de p-valores

Hess et al. [19] transformaram os dados de nível de expressão para a escala log_{10} para realizar a análise. Em seguida eles aplicaram o teste-t para variâncias diferentes e ordenaram os genes pelo ranking de p-valores.

O teste-t é um método estatístico muito utilizado para a detecção de genes diferencialmente expressos. Ele consiste em um teste de hipótese nula, verificando se as médias de duas distribuições normais são iguais. Para isso é calculado o p-valor como sendo a probabilidade, sobre a hipótese nula, de que o valor observado será tão ou mais elevado que o valor do teste estatístico calculado pela equação 4.1, que considera as amostras independentes e com variâncias desconhecidas e diferentes.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{4.1}$$

onde $\bar{x}$ é a média dos valores dos genes da classe 1, $\bar{y}$ é a média dos valores dos genes da classe 0, s_x e s_y são, respectivamente, o desvio padrão para a

classe 1 e o desvio padrão para a classe 0, n é o número de amostras de genes da classe 1 e m o número de amostras dos genes da classe 0.

Utilizando os genes mais informativos segundo o critério do ranking de p-valores os autores criaram previsores multigênicos utilizando os classificadores SVM, KNN e DLDA. Eles realizaram validação cruzada para cada combinação de genes e avaliaram a performance de cada classificador observando a AAC (Area Above the ROC Curve) e verificaram que o melhor classificador era composto pelas 30 sondas mais informativas segundo o ranking de p-valores, aplicadas ao classificador DLDA (Diagonal Linear Discriminant Analysis).

A tabela 4.1 mostra as 30 sondas selecionadas pelos autores [19].

Tabela 4.1: Sondas selecionadas por Hess et al. [19]

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

4.2.2 Seleção baseada em intervalos de níveis de expressão

Natowicz et al. [27] propuseram uma nova técnica para seleção de sondas que é bastante simples e intuitiva.

Considere que cada sonda s seja um classificador independente capaz de indicar se uma paciente é PCR ou NoPCR. Calcula-se a média e o desvio padrão do nível de expressão de cada sonda para o grupo PCR e o grupo NoPCR. Define-se dois intervalos de nível de expressão,intervalo PCR e intervalo NoPCR, para cada sonda como sendo [27]:

$$\mu(s)_{PCR} \pm \sigma(s)_{PCR} \tag{4.2}$$

$$\mu(s)_{NoPCR} \pm \sigma(s)_{NoPCR} \tag{4.3}$$

onde $\mu(s)_{PCR}$ e $\sigma(s)_{PCR}$ são, respectivamente, a média e o desvio padrão do nível de expressão de uma sonda *s* referente ao conjunto PCR e $\mu(s)_{NoPCR}$ e $\sigma(s)_{NoPCR}$ os valores referentes ao conjunto NoPCR.

Observando as equações 4.2 e 4.3 verifica-se que os intervalos podem ter interseção. Desta forma considera -se que todo valor de nível de expressão que esteja fora dos intervalos ou na interseção entre eles é considerado indefinido (UNDEFINED) [27]. As figuras 4.2.2 e 4.2.2 mostram essa definição.

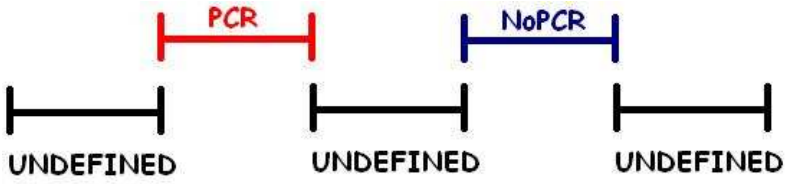


Figura 4.1: Intervalos PCR, NoPCR e UNDEFINED

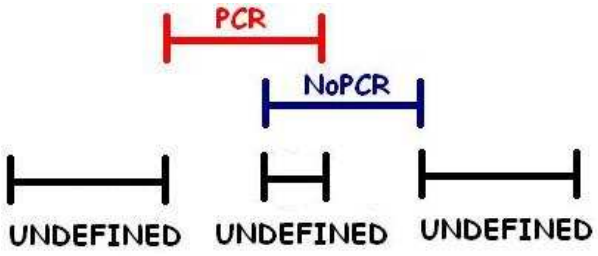


Figura 4.2: Intervalos PCR, NoPCR e UNDEFINED com interseção

Após calcular todos os intervalos PCR e NoPCR para todas as sondas utiliza-se o conjunto de treinamento para calcular o poder classificador de cada sonda [27] dado por:

$$V(s) = 0.5(\frac{p}{P} + \frac{n}{N}) \tag{4.4}$$

onde:

- p = número de pacientes PCR classificados corretamente
- P = número real de pacientes da classe PCR
- n = número de pacientes NoPCR classificados corretamente
- N = número real de pacientes da classe NoPCR

As sondas são ordenadas em ordem decrescente do valor de V(s).

Natowicz et al. [27] buscaram um conjunto de sondas que classificasse bem o conjunto de treinamento e o conjunto de teste. Para tanto utilizou o voto majoritário das sondas e verificou qual a quantidade de sondas necessária para obter essa classificação equilibrada, chegando a um conjunto de 30 sondas.

A tabela 4.2 mostra as 30 sondas selecionadas por Natowicz et al. [27].

Tabela 4.2: Sondas selecionadas por Natowicz et al. [27]

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

4.2.3 *Seleção baseada na técnica* Volcano plot

A técnica *Volcano Plot* consiste simplesmente em um gráfico que relaciona os resultados do *teste-t* e do *fold change* [12]. O teste-t utilizado foi o mesmo apresentado na seção 4.2.1 sendo que foram selecionados apenas os genes com p-valor menor que 0,05.

A forma mais simples para identificar quais genes são diferencialmente expressos é 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 [12]. Desta forma verifica-se 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. Essa técnica é chamada *fold change*. No caso desta dissertação verificamos se o valor de expressão do gene em uma das classes era 2 vezes maior ou menor que o valor de expressão do mesmo gene em outra classe [12].

Volcano plot é um *scatter-plot* de $-log_{10}$ dos p-valores calculados pelo teste-t para cada sonda, pelo *fold-change* calculado utilizando o log_2 da média das

razões entre as expressões dos genes de cada classe. A figura 4.3 mostra a técnica aplicada à base de dados de treinamento utilizada nessa dissertação.

Os genes mais diferencialmente expressos são aqueles que estão acima da linha tracejada horizontal e mais à esquerda da linha tracejada vertical esquerda (chamados *down regulated*) e mais a direita da linha tracejada vertical direita (chamados *up regulated*).

Utilizando essa técnica para analisar os dados de treinamento conseguiu-se obter as sondas mais significantes das classes PCR e NoPCR. Desta forma foram utilizados estes genes mais diferencialmente expressos para construir classificadores.

O método *volcano plot* selecionou 39 sondas no grupo *up regulated* (ou seja mais expressos na classe PCR) e 42 no grupo *down regulated* (mais expressos na classe NoPCR). Foi observado que entre as sondas selecionadas, algumas representavam o mesmo gene, fato devido à redundância presente na tecnologia Affymetrix, como visto anteriormente. Desta forma, apenas as sondas mais expressas foram mantidas, sendo que as demais que representavam o mesmo gene foram retiradas dos conjuntos. Fazendo isso restaram 31 sondas no grupo *up regulated* e 33 sondas no grupo *down regulated*.

Para selecionar o melhor conjunto de sondas classificadoras foram escolhidos subconjuntos de sondas *up regulated* e *down regulated* com o mesmo número cada um. O número variou de 1 a 31 sondas em cada grupo, possibilitando gerar classificadores que utilizam de 2 a 62 sondas. O motivo de ter se mantido igual o número de sondas de cada conjunto durante os teste foi evitar priorizar que características *up regulated* ou *down regulated* prevalecessem. Os classificadores escolhidos para utilizar os conjuntos de sondas foram *naïve* Bayes e o classificar de voto majoritário devido à simplicidade de ambos.

O conjunto de treinamento foi dividido em 10 partes para realizar uma validação cruzada. Desta forma o treinamento dos classificadores foi realizado 10 vezes, sendo que em cada vez uma das partes era guardada para ser utilizada como conjunto de validação. Os conjuntos foram montados de forma aleatória, sem a preocupação de manter as classes com a mesma proporção de PCR e NoPCR. Ao final dos 10 treinamentos da validação cruzada foram utilizados os resultados das classificações de cada conjunto de validação para montar uma classificação do conjunto original. A partir dessa classificação do conjunto original calculou-se a AAC (área acima da curva ROC: 1-AUC) para cada quantidade de sondas. Esse processo foi repetido 10 vezes para cada classificador e foi tomada a média das 10 AACs obtendo-se a Figura 4.4. A cada repetição os conjuntos de treinamento e validação foram refeitos de forma aleatória.

A Figura 4.4 mostra que a melhor classificação no conjunto de treinamento

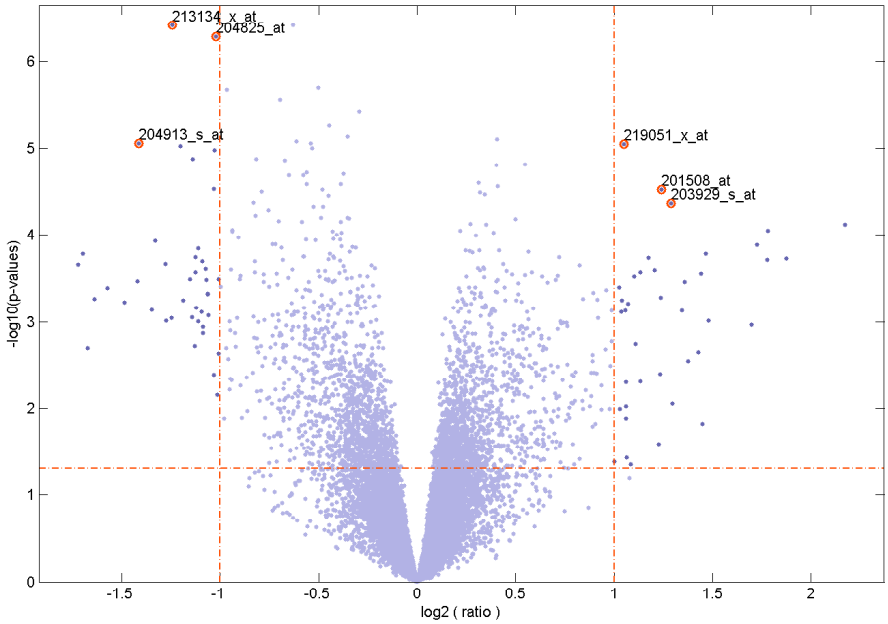


Figura 4.3: Volcanoplot

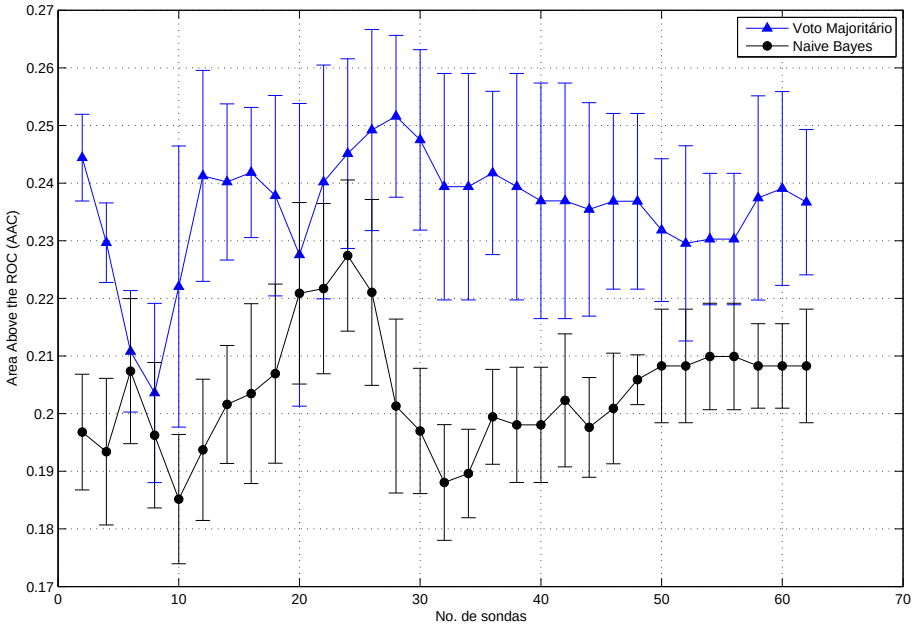


Figura 4.4: Área média acima da curva ROC

para o método de voto majoritário ocorreu para o subconjunto das 4 primeiras sondas de cada grupo, ou seja , as 4 primeiras sondas *up regulated* e as 4 primeiras *down regulated* resultando em um classificador de 8 sondas, que são representados pelas 4 primeiras linhas da tabela 4.3. Já para o classificador naïve Bayes a melhor classificação ocorreu para o conjunto de 10 sondas, sendo 5 *down regulated* e 5 *up regulated* que são representadas pelas

5 primeiras linhas da tabela 4.3. Nota-se, também, que uma boa classificação ocorreu para o conjunto de 32 sondas (16 de cada tipo) que são representadas por todas as linhas da tabela 4.3

Foi observado ainda que para o classificador bayesiano a AAC aumenta após as 10 primeiras sondas e volta a decrescer após 26 sondas até atingir o novo mínimo em 32 sondas. Desta forma foi selecionado um conjunto composto pelas 10 primeiras sondas e pelas sondas presentes no intervalo 26 a 32, já que elas tendem a melhorar o resultados. As demais sondas foram desprezadas por comprometerem a classificação. Assim foi obtido um conjunto com um total de 18 sondas que estão representadas na tabela 4.4.

No próximo capítulo serão apresentados os resultados para a classificação dos dados de treinamento e teste utilizando o conjunto de 30 sondas de Hess et al. [19], o conjunto de 30 sondas de Natowicz et al. [27], as 8 sondas que atingiram o mínimo de AAC para o método de voto majoritário, as 10 sondas e as 32 sondas presentes nos dois mínimos de AAC para o classificador bayesiano e por fim as 18 sondas presentes nos intervalos em que a AAC do classificador bayesiano diminuiu. Todos esses 6 conjuntos de sondas foram testados nos seguintes classificadores: voto majoritário, DLDA, *naïve* Bayes, SVM, comitê de perceptrons, rede neurais MLP, redes neurais com treinamento MOBJ e com treinamento LASSO.

Tabela 4.3: 32 sondas selecionadas utilizando *Volcano Plot*

Sonda *Up Regulated*	**Gene**	**Sonda *Down Regulated***	**Gene**
219051_x_at	METRN	205548_s_at	BTG3
201508_at	IGFBP4	204825_at	MELK
203929_s_at	MAPT	204913_s_at	SOX11
205225_at	ESR1	210147_at	ART3
214164_x_at	CA12	217028_at	CXCR4
212956_at	KIAA0882	213060_s_at	CHI3L2
218211_s_at	MLPH	204162_at	KNTC2
219197_s_at	SCUBE2	220625_s_at	ELF5
209173_at	AGR2	204533_at	CXCL10
212195_at	IL6ST	221872_at	RARRES1
209604_s_at	GATA3	202037_s_at	SFRP1
205696_s_at	GFRA1	209290_s_at	NFIB
217838_s_at	EVL	205044_at	GABRP
203628_at	IGF1R	208370_s_at	DSCR1
203789_s_at	SEMA3C	220559_at	EN1
221728_x_at	XIST	202342_s_at	TRIM2

Tabela 4.4: 18 sondas selecionadas utilizando a curva de AAC

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

4.3 Conclusões do capítulo

Este capítulo apresentou a base de dados utilizada na dissertação e três técnicas de seleção de sondas. Duas das técnicas foram apresentadas por importantes trabalhos da literatura, sendo que elas foram aplicadas sobre a mesma base de dados dessa dissertação. A terceira é a metodologia fundamental proposta neste trabalho, sendo baseada na técnica *Volcano Plot* e no classificador *naïve Bayes* com avaliação da performance pela AAC.

CAPÍTULO 5

Classificadores Utilizados

O problema de previsão de PCR consiste em classificar os pacientes em sensíveis ao tratamento (PCR) ou não (NoPCR). Este problema de classificação é muito complicado pois as classes são desbalanceadas, sendo que aproximadamente 70% dos padrões são da classe negativa (NoPCR) e o restante da classe positiva (PCR). Outro agravante é a escassez de dados o que dificulta ainda mais a obtenção de um modelo que generalize bem. Buscando contornar estes problemas foram utilizados os classificadores apresentados nesse capítulo sobre cada um dos conjuntos de sondas selecionados.

Assim este capítulo apresenta uma breve introdução sobre cada um dos classificadores utilizados.

5.1 Voto Majoritário

O classificador por voto majoritário foi proposto por Natowicz et al. [27]. Ele consiste em verificar se o valor de cada sonda selecionada pertence ao intervalo PCR ou No-PCR ou se é UNDEFINED como descrito na sessão 4.2.2. Desta forma conta-se quantas sondas são da classe PCR e quantas são da classe No-PCR, sendo a classe do paciente aquela que tiver maior número de votos. Em caso de empate o paciente é considerado PCR.

5.2 Classificador Naïve Bayes

O teorema de Bayes diz que:

$$p(C|x_1, ..., x_n) = \frac{p(C)p(x_1, ..., x_n|C)}{p(x_1, ..., x_n)}. \tag{5.1}$$

O classificador naïve Bayes faz a consideração "ingênua" de que todas as variáveis de entrada são independentes, ou seja [5]:

$$p(x_1, ..., x_n|C) = P(x_1|C) \cdot ... \cdot P(x_n|C) \tag{5.2}$$

Dessa forma, a partir da equação 5.1, pode-se chegar à seguinte regra de decisão para um problema de duas classes [15]:

$$Classe(x_1, ..., x_2) = \begin{cases} C_1 & \text{se } \frac{p(x_1,...,x_n|C_1)}{p(x_1,...,x_n|C_1)} \leq k \\ C_2 & \text{caso contrário} \end{cases} \tag{5.3}$$

onde k = $\frac{p(C_2)}{p(C_1)}$. Fazendo a consideração "ingênua" de independência, a equação 5.3 se reduz a equação 5.4:

$$Classe(x_1, ..., x_2) = \begin{cases} C_1 & \text{se } \frac{p(x_1|C_1)\cdot...\cdot p(x_n|C_1)}{p(x_1|C_2)\cdot...\cdot p(x_n|C_2)} \leq k \\ C_2 & \text{caso contrário} \end{cases} \tag{5.4}$$

A desvantagem desse classificador é a consideração de independência entre as variáveis, uma vez que nem sempre essa consideração é válida. Contudo, esse classificador funciona bem para vários problemas práticos.´

É necessário também especificar um modelo probabilístico para $p(x_i, ..., x_n|C_j)$. Nesta dissertação foi utilizado o modelo de distribuição probabilístico normal.

5.3 Classificador DLDA

O classificador DLDA (Diagonal Linear Discriminant Analysis) é uma regra de máxima verossimilhança para problemas de distribuição normal multivariada onde as distribuições das classes têm a mesma matriz diagonal de variância-covariância, ou seja as variáveis são não-correlacionadas e para cada variável a variância é a mesma em todas as classes [35].

O padrão x será pertencente à classe 1 se

$$\sum_{i=1}^{S} \frac{(x_i - \bar{x_i}^{(1)})^2}{\sigma_i^2} \leq \sum_{i=1}^{S} \frac{(x_i - \bar{x_i}^{(2)})^2}{\sigma_i^2} \tag{5.5}$$

caso contrário será pertencente à classe 2. Na equação 5.5 S é o número de variáveis que formam o padrão, $\bar{x_i}^{(1)}$ é o valor médio da variável i para a classe 1 e $\bar{x_i}^{(2)}$ é o valor médio da variável i para a classe 2 e σ_i^2 é a variância da variável i sobre o conjunto completo, formado pelas duas classes.

5.4 SVM

SVM (*Support Vector Machine*) é um modelo de aprendizado de máquina muito utilizado para resolver problemas de classificação. É baseado na minimização do risco estrutural visando a construir um conjunto de hiperplanos variando a dimensão VC (dimensão Vapnik e Chervonenkis), de modo que o erro de treinamento e a dimensão VC sejam minimizados ao mesmo tempo [41, 34]. Dessa forma busca-se o hiperplano ótimo que maximize a margem de separação entre as classes. Encontrar este hiperplano consiste em resolver o problema dual ao problema de minimização da dimensão VC. Minimizar a dimensão VC é o mesmo que maximizar a margem de separação [41, 34].

Para um problema de classificação binária onde (x_i,y_i) forma o par entrada-saída para cada padrão de treinamento e $y_i \in \{+1, -1\}$ temos que o problema de otimização dual para encontrar o hiperplano ótimo é dado por:

$$Maximizar: \quad W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{5.6}$$

$$Sujeito \quad a: \quad \sum_{i=1}^{N} y_i \alpha_i = 0; \forall_{i=1}^{N} : 0 \leq \alpha_i \leq C. \tag{5.7}$$

onde $W(\alpha)$ é o vetor de pesos, o parâmetro C é especificado pelo usuário, α_i é o multiplicador de Lagrange, $K(x_i, x_j)$ é o kernel utilizado. O parâmetro C controla a relação entre a complexidade do algoritmo e o número de amostras de treinamento classificados incorretamente. Ele pode ser visto como um parâmetro de penalização [34].

O vetor de pesos ótimo w^* é dado por:

$$w^* = \sum_{i=1}^{N} y_i \alpha_i^* x_i \tag{5.8}$$

e o intercepto b^* é dado por:

$$b^* = -\frac{1}{2} \left[\max_{\{i|y_i=-1\}} \left(\sum_{j=1}^{N_{SV}} y_j \alpha_j K(x_i, x_j) \right) + \min_{\{i|y_i=+1\}} \left(\sum_{j=1}^{N_{SV}} y_j \alpha_j K(x_i, x_j) \right) \right] \tag{5.9}$$

A função de decisão é dada por [34]:

$$f(x) = sgn \left(\sum_{i=1}^{N_{SV}} y_i \alpha_i^* K(x_i, x) + b^* \right) \tag{5.10}$$

ou seja, se $f(x) \geq 0$ então x pertence a classe +1 senão x pertence a classe -1. Nessa equação N_{sv} é o número de vetores de suporte.

Para resolver problemas não-linearmente separáveis é necessário fazer um mapeamento não-linear para um espaço característico onde os dados serão linearmente separáveis. Para tanto utiliza-se uma função de mapeamento $\phi(x)$.

A superfície de separação no espaço característico será dado por [34]:

$$\sum_{i=1}^{N} \alpha_i y_i \phi(x)^t \phi(x) = 0 \tag{5.11}$$

O produto interno $\phi(x)^t\phi(x)$ é chamado produto interno *Kernel* ou $K(x, x_i)$. O que diferencia um modelo de SVM de outro é exatamente o tipo de kernel utilizado, ou seja, cada modelo de SVM utiliza um mapeamento diferente.

Os tipos de *kernel* utilizados nessa dissertação são:

- **Kernel linear:**

$$K(x, x) = x \cdot x^t \tag{5.12}$$

  onde x é o padrão de entrada.

- **Kernel RBF**

$$K(x, x_i) = exp(-\left\|x_i - x\right\|^2 / 2\sigma^2) \tag{5.13}$$

  onde o parâmetro σ^2 é especificado a priori pelo usuário.

5.5 Comitê de Perceptrons

Os neurônios artificiais ou *perceptrons* são compostos por entradas, cada uma com um peso específico, função de ativação que é aplicada ao resultado da soma ponderada das entradas e saída que emite o resultado produzido pelo neurônio.

Um comitê de *perceptrons* consiste simplesmente de vários neurônios artificiais treinados independentemente com os mesmos dados de treinamento. A classe dos padrões de entrada é a que aparecer mais vezes no comitê.

5.6 Redes Neurais Multi-Layer Perceptron (MLP)

Uma rede multicamada é composta por uma camada de entrada, uma ou mais camadas escondidas e uma camada de saída. A rede pode ser completamente conectada onde cada neurônio está ligado a todos os neurônios da camada seguinte, parcialmente conectada onde cada neurônio está ligado a um certo número de neurônios da camada seguinte, ou localmente conectada onde há uma conexão parcial orientada para cada tipo de funcionalidade. A Figura 5.1 mostra um exemplo de rede MLP.

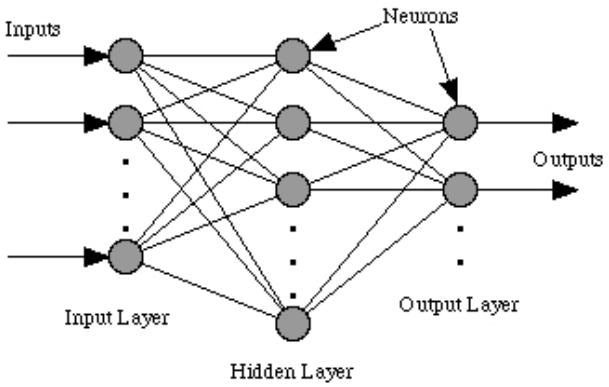


Figura 5.1: Arquitetura feedforward

Para realizar o treinamento de uma rede neural é necessário um conjunto de dados que contenha padrões para o treinamento e saídas desejadas. Desta forma o problema do treinamento das redes neurais se resume em um problema de otimização em que se deseja encontrar o melhor conjunto de pesos que minimize o erro médio quadrático calculado entre as saídas da rede e as saídas desejadas. A equação 5.14 mostra como é realizado o cálculo.

$$MSE = \frac{1}{N_T} \sum_{i}^{N_T} (yd_i - ys_i)^2 \tag{5.14}$$

Nessa equação temos que yd_i representa as saídas desejadas para cada padrão de treinamento e ys_i representa a saída simulada pela rede para cada padrão apresentado. N_T é o número de padrões do conjunto de treinamento.

O algoritmo mais utilizado para solucionar este problema é o *backpropagation*, que é um algoritmo baseado em direção de busca e que propaga o erro das camadas finais da rede para as iniciais [8].

5.7 MOBJ-NN

Uma questão importante no treinamento de redes neurais é a capacidade de generalização da rede resultante. Muitas vezes obtém-se redes com baixo erro médio quadrático, mas que não generalizam bem, ficando super-ajustadas aos padrões de treinamento. Outras vezes as redes obtidas ficam sub-ajustadas aos padrões de treinamento, não aprendendo corretamente a tarefa ensinada.

Boa capacidade de generalização é alcançada se alguns fatores são levados em consideração, como por exemplo, o número de padrões utilizados no processo de treinamento e a capacidade do modelo de se ajustar a estes dados. Considerando que o tamanho do conjunto de treinamento é adequado para uma dada tarefa de aprendizagem, ou seja, é estatísticamente representativo, a capacidade de generalização pode ser conseguida se o modelo tiver complexi-

dade ótima. Modelos muito complexos (os quais se ajustam bem aos dados) e modelos pouco complexos (os quais se ajustam mal aos dados) freqüentemente apresentam baixa capacidade de generalização. Desta forma deve haver um equilíbrio entre a complexidade do modelo e a complexidade inerente à tarefa [39].

Na literatura existem duas formas principais para controlar a complexidade da rede, a primeira e mais comum diz respeito ao número de neurônios da camada escondida. Essa abordagem se sub-divide em duas, sendo uma referente a minimização do número de neurônios da camada escondida (abordagem mais utililizada com algoritmos evolucionários) e outra referente à poda de neurônios pouco expressivos. A segunda forma consiste em controlar a magnitude dos pesos da rede neural, minimizando a norma euclidiana do vetor composto por todos os pesos da rede [39].

O algoritmo MOBJ [39] utiliza a minimização da norma euclidiana dos pesos juntamente com a minimização do erro médio quadrático para treinar redes controlando sua complexidade. O algoritmo utiliza a técnica de otimização multiobjetivo ϵ-restrito (P_ϵ), tomando como restrição a norma euclidiana dos pesos e minimizando o erro médio quadrático.

Para resolver o problema P_ϵ proposto pode ser utilizado qualquer algoritmo para otimização com restrições. Na versão original [39] é utilizado o algoritmo elipsoidal [6] e nessa dissertação será utilizado o algoritmo de gradiente projetado da forma apresentada por Horta et al. [20].

5.7.1 Otimização Multiobjetivo

Na otimização multiobjetivo deseja-se otimizar duas ou mais funções objetivo ao mesmo tempo. Durante este processo podem surgir dois tipos de soluções [38]:

- Soluções que, sob todos os objetivos simultaneamente considerados, serão dominadas por outras soluções. Isso significa que há soluções que conseguem valores melhores de função objetivo para todos os objetivos considerados. Essas soluções devem ser descartadas.

- Soluções que, comparadas com outras soluções, serão melhores em algum ou alguns objetivos, mas piores em outro ou outros objetivos. Essas soluções não devem ser descartadas, devendo um decisor escolher aquela que melhor se adequa a uma determinada característica do problema.

O segundo grupo de soluções consiste no conjunto de soluções eficientes ou *Pareto-ótimas*.

Desta forma, a solução de um problema de otimização multiobjetivo será um conjunto de soluções ao invés de uma solução única.

Dominância

Para um problema de minimização de duas funções objetivo diz-se que um ponto x_1, pertencente a um conjunto de soluções, domina o ponto x_2, também pertencente ao conjunto de soluções, se x_1 for melhor em ao menos uma função objetivo e não for pior em nenhuma das outras. Diz-se que x, pertecente ao conjunto factível, é uma solução Pareto-ótima se não existe qualquer outra solução x, pertecente ao conjunto factível, que a domine.

Se existisse uma única solução no conjunto factível que dominasse todas as demais ela seria a solução utópica. Esta solução recebe este nome porque não pode ser alcançada em um problema de otimização multiobjetivo, caso ela fosse alcançada o problema se tornaria mono-objetivo.

5.7.2 Método MOBJ para treinamento de redes neurais

O método de treinamento MOBJ consiste em controlar a complexidade das redes através da minimização simultânea do erro para os padrões de treinamento e da norma do vetor de pesos. A minimização destes objetivos é feita até que um ponto de equilíbrio seja alcançado, obtendo-se as soluções chamadas de não-dominadas ou Pareto-ótimas. Estas soluções são aquelas as quais não há mais como melhorar um dos objetivos sem que haja uma degradação do outro, ou seja, considerando um determinado erro ϵ, a norma correspondente a este é a menor norma possível. Por outro lado, considerando uma determinada norma N, o erro correspondente a esta é o menor erro possível [39].

Este algoritmo tem duas versões, sendo que a primeira usa o método ϵ-restrito enquanto a segunda utiliza uma variante do primeiro, denominada método das relaxações. A abordagem testada neste trabalho foi apenas a ϵ-restrito.

O algoritmo MOBJ obtém um determinado número ς de soluções a ser informado pelo usuário com diferentes complexidades. Cada uma destas soluções é uma solução não-dominada constituindo o conjunto Pareto-ótimo. A formulação do método ϵ-restrito (P_ϵ) aplicado ao problema de treinamento de redes neurais artificiais é feita a seguir [39]:

$$\min_{w \in W} f_1(w)$$

sujeito a :

$$f_2(w) \leq \epsilon \tag{5.15}$$

onde:

$$f_1(w) = \frac{1}{N_T} \sum_{i}^{N_T} (y_i - \hat{y}_i(w))^2 \tag{5.16}$$

$$f_2(w) = \|\mathbf{W}\| \tag{5.17}$$

Nas equações w representa o vetor de pesos da rede, y_i é a saída desejada para o padrão de treinamento i, $\hat{y}_i(w)$ é a saída simulada para o mesmo padrão e N_T é o número de padrões de treinamento. A Figura 5.2 mostra como é feita a geração de soluções via P_ϵ. O algoritmo 1 mostra a implementação do MOBJ utilizando o método P_ϵ.

Algoritmo 1: MOBJ - Obtenção do conjunto Pareto-ótimo - Método P_ϵ

```
Carregar conjunto de treinamento;
N_N ⇐ Número de neurônios da camada intermediária;
ς ⇐ Número de soluções a serem geradas;
δ_ε ⇐ Diferença entre as normas das soluções obtidas;
Inicializar ε;
// ε = δ_ε
Inicialiar w_0;
/* Pesos aleatórios, média zero e variância pequena       */
C_po ⇐ 1;
// Contador de soluções Pareto
while C_po ≤ ς do
    Obter w* que minimize f_1(w) sujeito a:
    Restrição 1: g_1(w) = ||w|| ≤ ε;
    /* Utilizar algum algoritmo de otimização com restrições
       */
    w_0 ⇐ w*;
    /* A próxima busca se iniciará no ponto ótimo w* da busca
       anterior                                               */
    Guardar solução w*;
    /* w* é uma solução do conjunto Pareto-ótimo           */
    ε = ε + δ_ε;
    C_po = C_po + 1;
end
```

Na versão original [39] o algoritmo utilizado para encontrar os vetores de pesos do Pareto foi o Elipsoidal [6]. Neste trabalho será testado outro algoritmo, o Gradiente Projetado [29, 30]. Os dois algoritmos são apresentados nas seções a seguir.

5.7.3 Algoritmo Elipsoidal

O algoritmo elipsoidal é um algoritmo de exclusão de semi-espaços. Ele surgiu na década de 70 e foi proclamado como o substituto do *Simplex*, o que

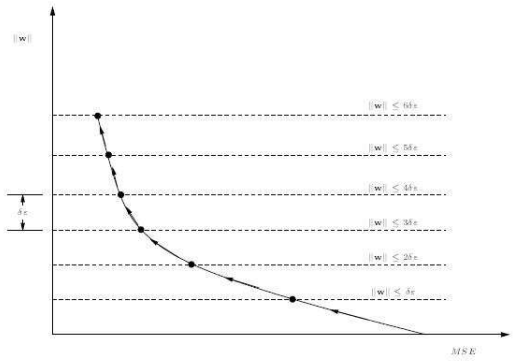


Figura 5.2: Geração de soluções via P_ϵ

nunca aconteceu.

A idéia intuitiva do algoritmo é cercar a solução ótima com um elipsóide. A partir desta elipsóide são feitos cortes, gerando sempre elipsóides menores, até que se chegue a um elipsóide degenerado sobre o ponto ótimo, que será seu centro.

O algoritmo elipsoidal básico é descrito pelas seguintes fórmulas recursivas que geram uma seqüência de pontos x_k [38, 6]:

$$w_{k+1} = w_k - \beta_1 \frac{Q_k g_k}{(g_k Q_k g_k)^{1/2}} \tag{5.18}$$

$$Q_{k+1} = \beta_2 \left(Q_k - \frac{\beta_3 (Q_k g_k)(Q_k g_k)^T}{g_k^T Q_k g_k} \right) \tag{5.19}$$

com

$$\beta_1 = \frac{1}{n+1} \quad \beta_2 = \frac{n^2}{n^2-1} \quad \beta_3 = \frac{2}{n+1} \tag{5.20}$$

O vetor g_k é o subgradiente da restrição mais violada $f_i(w_k)$, $i \geq 1$, ou, no caso de w_k estar na região factível, um subgradiente da função objetivo $f_0(w_k)$ naquele ponto [38].

Uma forma para se cortar a elipsóide a fim de se utilizar menor número de iterações do algoritmo é apresentada a seguir.

5.7.4 *Algoritmo Elipsoidal com* Deep Cut

O procedimento convencional de *deep cut* é descrito pelas equações 5.18 e 5.19 onde os valores de β_1, β_2 e β_3 são calculados como a seguir [31, 38]:

$$\beta_1 = \frac{1 + n\alpha}{n+1} \tag{5.21}$$

$$\beta_2 = \frac{n^2(1-\alpha^2)}{n^2-1} \tag{5.22}$$

$$\beta_3 = \frac{2(1-n\alpha)}{(n+1)(1+\alpha)} \tag{5.23}$$

5.7.5 Método de Gradiente Projetado

O método Gradiente Projetado foi proposto em 1960 e sua descrição pode ser vista detalhadamente em [29, 30, 3]. A idéia é projetar o negativo de gradiente em um subespaço factível, de forma a otimizar a função objetivo e manter as condições de factibilidade.

O vetor w_k é atualizado da seguinte forma:

$$w_k = w_k - \alpha d_k \tag{5.24}$$

onde α é o argumento que minimiza $f(w_k - \alpha d_k)$, d_k = $-P\nabla f(x)$ e P é a matriz de projeção definida por:

$$P = I - M^t(MM^t)^{-1}M \tag{5.25}$$

sendo M uma matriz em que as linhas são os gradientes das restrições.

Se $d_k \neq 0$ então minimize $f(w)$ começando em w_k na direção d_k e faça um movimento de correção para a região factível. Se, caso contrário, $d_k = 0$ então calcule (u^t,v^t) = $-\nabla f(w)^t M^t(MM^t)^{-1}$. Se $u \geq \mathbf{0}$ então pare com um ponto w_k que satisfaça as condições de Kuhn-Tucker. Caso contrário, delete a linha de M correspondente a algum $u_i < 0$ e repita o processo [3].

A correção (corr) para atingir a região factível é feita com base na violação das restrições ativas, da seguinte forma:

$$corr = -M^t(MM^t)^{-1}g(w_k) \tag{5.26}$$

onde $g(w_k)$ é o valor da restrição violada no ponto w_k. Assim o valor de w_{k+1} é ajustado da seguinte forma:

$$w_k = w_k - \alpha d_k + corr \tag{5.27}$$

A direção d_k é normalizada em relação ao maior valor absoluto presente no próprio vetor.

Um dos passos mais importantes deste algoritmo é o cálculo do passo α que é dado na direção de gradiente projetado. Para calcula-lo expandimos $f(w_k)$ em série de Taylor de primeira ordem:

$$f(w_{k+1}) = f(w_k) + \Delta w^T \nabla f(w_k) \tag{5.28}$$

Definindo que $f(w_{k+1})$ seja apenas 5% menor que o valor anterior temos que $f(w_{k+1})$ = $0.95f(w_k)$. Sabemos ainda que:

$$w_(k+1 = w_k - \alpha d_k \tag{5.29}$$

onde d_k é a direção de busca. Assim $\Delta w = w_{k+1} - w_k = -\alpha d_k$. Substituindo em 5.28 temos:

$$f(w_{k+1}) - f(w_k) = -\alpha d_k \nabla f(w_k) \tag{5.30}$$
$$0.95f(w_k) - f(w_k) = -\alpha d_k \nabla f(w_k) \tag{5.31}$$
$$0.05f(w_k) = \alpha d_k \nabla f(w_k) \tag{5.32}$$
$$\alpha = \frac{0.05f(w_k)}{d_k \nabla f(w_k)} \tag{5.33}$$

A partir da equação 5.33 e considerando uma modificação menor no valor de $f(w_k + 1)$ em relação a $f(w_k)$ no início das iterações desenvolvemoso algoritmo 2 onde $epslon$ é o menor valor considerado como sendo diferente de zero. Neste algoritmo MAXITER é o número máximo de iterações para o método de gradiente projetado. Desta forma o parâmetro α caminha em frações menores da direção de busca a cada iteração, fazendo uma busca suave.

Algoritmo 2: Cálculo do parâmetro α

```
// Cálculo do alfa ótimo inicial.
if ||d_k^t * ∇f(w_k)|| > epslon then
    α = |(0.05 − 0.025(k+1)/MAXITER) * f(w_k)/(d_k^t * ∇f(w_k))|;
end
// Verifica se a norma da direção de busca é diferente de
   zero
if ||d_k|| > epslon then
    if k > 1 then
        α = (0.05 − 0.025(k+1)/MAXITER);
        if ||d_k^t * ∇f(w_k)|| > epslon then
            α = |α f(w_k)/(d_k^t * ∇f(w_k))|;
        end
    end
end
```

5.8 LASSO

O treinamento multiobjetivo de redes neurais artificiais pelo método LASSO (*least absolute shrinkage and selection operator*)[40] consiste no mesmo algo-

ritmo do MOBJ-NN, com a diferença na restrição de norma dos pesos, sendo que neste a norma é a L_2 e naquela é a norma é L_1 [10]. Assim o algoritmo se resume a:

$$\begin{gathered}\mathbf{w}^* = arg\ \min \frac{1}{N}\sum_{j=1}^{N}\left(d_j - y(\mathbf{w},\mathbf{x}_j)\right)^2 \\ sujeito\,a:\ \sum_i |w_i| \leq \epsilon\end{gathered} \tag{5.34}$$

onde w representa o vetor composto por todos os pesos da rede, d_j é a solução desejada para cada padrão, $y(\mathbf{w},\mathbf{x}_j)$ é a saída calculada pela rede para cada padrão, N é o número de padrões de treinamento e ϵ é a restrição de norma em cada passo do algoritmo P_ϵ.

Para realizar a otimização dos pesos é utilizado o algoritmo elipsoidal [6], assim como na versão original do MOBJ-NN.

Essa escolha de um tipo de norma diferente para o controle de complexidade da rede tem um impacto grande na topologia final, já que cada vez que o algoritmo de otimização encontra a solução em um vértice da restrição alguns dos pesos da rede ficam iguais a zero, sendo que assim o treinamento multiobjetivo LASSO faz *pruning* automático da rede reduzindo não só o número de pesos utilizados, como também o número de entradas, selecionando assim aquelas entradas que são realmente relevantes para o aprendizado da rede.

5.9 Conclusões do capítulo

Neste capítulo foram apresentados cada um dos classificadores utilizados nessa dissertação. Em especial foi utilizado o algoritmo de gradiente projetado para o ajuste dos pesos da rede no algoritmo MOBJ assim como apresentado por Horta et al. [20].

CAPÍTULO 6

Resultados

Este capítulo apresenta os resultados da classificação dos pacientes em PCR e NoPCR para cada conjunto de sondas selecionado nesta dissertação aplicado a cada um dos classificadores escolhidos. Os resultados são apresentados em termos da acurácia (Ac), sensibilidade (Se) e especificidade (Es). Ao final do capítulo é feita a análise geral dos resultados em que é apresentado o melhor conjunto de sondas. É apresentado ainda um refinamento do conjunto escolhido, realizado pelo método LASSO para treinamento multiobjetivo de redes neurais.

6.1 Voto Majoritário

O classificador de voto majoritário [27] foi utilizado com cada um dos conjuntos de sondas selecionadas. Como era de se esperar a melhor solução foi aquela encontrada com o conjunto selecionado por Natowicz et al. [27], já que esse conjunto foi selecionado com base nesse classificador. A Tabela 6.1 mostra a acurácia (Ac), sensibilidade (Se) e especificidade (Es) para cada conjunto de sondas.

Tabela 6.1: Resultados para o classificador de voto majoritário das sondas

Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.84146	0.87805	0.79268	0.78049	0.78049	0.78049
Se	0.80952	0.7619	0.80952	0.80952	0.7619	0.7619
Es	0.85246	0.91803	0.78689	0.77049	0.78689	0.78689
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.86275	0.82353	0.80392	0.80392	0.86275	0.86275
Se	0.92308	0.46154	0.76923	0.84615	0.84615	0.76923
Es	0.84211	0.94737	0.81579	0.78947	0.86842	0.89474

6.2 DLDA

O classificador DLDA foi utilizado na forma descrita por Hess et al. [19], ou seja, os dados foram transformados para log_{10} a fim de reduzir a distância entre os dados. Como pode ser observado na Tabela 6.2 a melhor solução para este classificador foi o modelo que utiliza as 10 sondas selecionadas pelo método *Volcano Plot*, já que foi a solução mais equilibrada em termos de especificidade e sensibilidade tanto para o conjunto de treinamento quanto para o de validação.

Tabela 6.2: Resultados para o classificador DLDA

Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.82927	0.8171	0.78049	0.79268	0.8171	0.79268
Se	0.85714	0.9524	0.85714	0.80952	0.7619	0.80952
Es	0.81967	0.7705	0.7541	0.78689	0.8361	0.78689
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.84314	0.76471	0.84314	0.86275	0.8431	0.86275
Se	0.92308	0.92308	0.84615	0.84615	0.6923	0.76923
Es	0.81579	0.71053	0.84211	0.86842	0.8947	0.89474

6.3 Naïve Bayes

Todos os conjuntos de sondas foram testados no classificador *naïve* Bayes.Como pode ser observado todas as soluções, com a exceção do conjunto de sondas de Hess et al. [19] , foram bem equilibradas, com destaque para a solução obtida pelo conjunto de 18 sondas.

Tabela 6.3: Resultados para o classificador bayesiano

Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.82927	0.82927	0.81707	0.82927	0.84146	0.87805
Se	0.80952	1	0.7619	0.80952	0.76190	0.80952
Es	0.83607	0.77049	0.83607	0.83607	0.86885	0.90164
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.86275	0.82353	0.80392	0.84314	0.82353	0.86275
Se	0.76923	0.53846	0.69231	0.76923	0.76923	0.84615
Es	0.89474	0.92105	0.84211	0.86842	0.84211	0.86842

6.4 SVM

As SVMs necessitam que alguns parâmetros sejam ajustados para que uma boa solução seja calculada. Para a SVM com kernel linear é necessário o ajuste do parâmetro C e para a SVM com kernel RBF é necessário o ajuste dos parâmetros C e σ. Para tanto foi utilizada a metodologia semelhante à sugerida por Gestel et al. [18]. O conjunto de treinamento foi dividido em 10 conjuntos para realizar a validação cruzada. A cada vez são utilizados 9 conjuntos para treinar e 1 conjunto para validar, alterando o conjunto de validação a cada passo da validação cruzada. São definidos intervalos de valores para C e σ (este último apenas para o kernel RBF) e é feita a validação cruzada para cada um deles. Os parâmetros que derem os melhores resultados em termos da AUC (*Area Under the ROC*) são escolhidos e um novo intervalo de valores é definido em torno deles, de forma mais precisa. Esse processo é repetido 3 vezes, sendo os parâmetros escolhidos ao final deste processo. De posse do valor desses parâmetros, uma nova SVM é treinada utilizando todo o conjunto de treinamento. O intervalo de valores de C utilizado foi de 0.01 a 1000 e de sigma foi de 0.5 a 500.

6.4.1 SVM com kernel linear

A Tabela 6.4 mostra os parâmetros utilizados por cada SVM e os resultados para cada conjunto de sondas utilizado. Como pode ser observado os resultados foram bem equilibrados para os conjuntos de 10 e 18 sondas, tanto em treinamento quanto em teste. Já as sondas do modelo de Natowicz et al. [27] não tiveram bons resultados em termos do conjunto de teste. Pode ser observado ainda que o valor de C escolhido para o conjunto de 10 sondas é muito maior que aquele utilizado para o conjunto de 18 sondas.

Tabela 6.4: Resultados para SVM com kernel linear

Parâmetros						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
C	0.01	0.01	1	10	0.01	0.01
Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.87805	0.84146	0.86585	0.87805	0.86585	0.84146
Se	0.85714	0.95238	0.85714	0.85714	0.80952	0.80952
Es	0.88525	0.80328	0.86885	0.88525	0.88525	0.85246
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.72549	0.78431	0.80392	0.84314	0.84314	0.86275
Se	0.92308	0.92308	0.69231	0.84615	0.76923	0.84615
Es	0.65789	0.73684	0.84211	0.84211	0.86842	0.86842

6.4.2 SVM com kernel RBF

A Tabela 6.5 mostra os parâmetros C e σ utilizados por cada SVM com kernel RBF e os resultados para cada conjunto de sondas utilizado. Pode ser observado que as sondas selecionadas por Natowicz et al. [27] e Hess et al. [19] tendem a sofrer um super-ajuste aos dados de treinamento (*overfitting*). Já os demais conjuntos fornecem boas soluções, com destaque para os conjuntos de 32 e 18 sondas que foram os melhores classificadores, já que foram excelentes no conjunto de treinamento e muito bons no conjunto de teste, o que indica que generalizaram bem, sem sofrer com o *overfitting*.

Tabela 6.5: Resultados para SVM com kernel RBF

Parâmetros						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
C	10	5	1	2.25	0.05	0.3
σ	3	2.5	1.75	5.75	2.5	1.9
Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	1.000	1.000	0.878	0.841	0.939	0.951
Se	1.000	1.000	0.905	0.810	0.857	0.857
Es	1.000	1.000	0.869	0.852	0.967	0.984
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.686	0.784	0.824	0.843	0.863	0.863
Se	0.308	0.308	0.769	0.846	0.923	0.846
Es	0.816	0.947	0.842	0.842	0.842	0.868

6.5 Comitê de Perceptrons

O comitê foi formado por 11 *perceptrons* treinados com todo o conjunto de treinamento. O conjunto de treinamento foi normalizado usando a equação 6.1:

$$Pn = \frac{P - m}{DP} \tag{6.1}$$

onde Pn é o valor da sonda normalizada, P é o valor de nível de expressão da sonda a ser normalizada, m é a média dos níveis de expressão para a sonda em questão e DP é o desvio padrão dos níveis de expressão. Essa normalização faz com que as distribuições sejam todas com média zero e desvio padrão igual a um;

Foi escolhido um número ímpar de *perceptrons* para evitar empates. Essa abordagem força uma separação linear. Como pode ser observado os melhores resultados apareceram para as sondas selecionadas por Natowicz e as 32 sondas selecionadas pelo método de *volcano plot*.

Tabela 6.6: Resultados para o comitê de perceptrons

Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.927	0.854	0.854	0.841	0.878	0.939
Se	0.857	0.905	0.810	0.667	0.619	0.810
Es	0.951	0.836	0.869	0.902	0.967	0.984
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.824	0.706	0.784	0.804	0.824	0.804
Se	0.846	0.769	0.615	0.615	0.769	0.769
Es	0.816	0.684	0.842	0.868	0.842	0.816

6.6 Rede Neurais MLP

Para todos os conjuntos de sondas foi utilizada a mesma topologia, a saber: uma camada escondida, 5 neurônios na camada escondida, todos os neurônios utilizam a função de transferência tangente hiperbólica sigmoidal, foram utilizadas no máximo 300 épocas e o algoritmo de treinamento foi o Levenberg-Marquardt backpropagation. O treinamento foi executado 10 vezes utilizando todo o conjunto de treinamento normalizado pela equação 6.1. A solução escolhida foi a que teve o melhor resultado em teste entre as execuções.

A Tabela 6.7 mostra os resultados para cada conjunto de sondas. Como pode ser observado os melhores resultados ocorreram para os conjuntos de 32 e 18 sondas.

Tabela 6.7: Resultados para rede neural MLP

Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.890	0.793	0.890	0.927	0.878	0.902
Se	0.762	0.857	0.762	0.714	0.857	0.762
Es	0.934	0.770	0.934	1.000	0.885	0.951
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.667	0.824	0.784	0.824	0.863	0.843
Se	0.692	0.769	0.538	0.615	0.769	0.846
Es	0.658	0.842	0.868	0.895	0.895	0.842

6.7 MOBJ-NN

Para o modelo MOBJ-NN foi utilizada uma rede com 10 neurônios na camada escondida, norma máxima dos pesos igual a 4 e o passo da norma igual a 0.05, gerando um total de 80 soluções a cada execução. Todos os neurônios utilizam a função de transferência tangente hiperbólica. Os dados foram normalizados de acordo com a equação 6.1. O algoritmo de otimização utilizado para o ajuste dos pesos foi o algoritmo de gradiente projetado [29]. Durante a execução algumas soluções foram dominadas, por isso foi executado um teste de dominância para filtrar o Pareto, permitindo que apenas as soluções não dominadas fossem utilizadas.

O treinamento foi baseado em validação cruzada, sendo utilizados 70% dos dados de treinamento para treinar a rede e os 30% restantes foram utilizados para validar. Esses dois conjuntos foram montados de forma aleatória, mas mantendo a relação entre o número de pacientes PCR e NoPCR constante. O processo foi executado 10 vezes, obtendo assim 10 paretos. A solução es-

colhida em cada pareto foi aquela que teve maior AUC sobre o conjunto de validação. Em caso de empate a norma da solução escolhida foi a média das normas dos pesos das soluções que empataram. Feito isso tomou-se as médias das 10 normas obtidas, obtendo-se assim a norma desejada para a rede a ser escolhida.

O algoritmo foi executado mais 10 vezes, porém utilizando-se 100% do conjunto de treinamento. A solução escolhida em cada execução foi aquela cuja a norma dos pesos mais se aproximou da norma desejada calculada anteriormente. A solução final foi aquela que mais se repetiu durante as ultimas 10 execuções do algoritmo.

Os resultados para o treinamento do MOBJ-NN com o critério de escolha descrito encontram-se na Tabela 6.8.

Como pode ser observado os melhores resultados foram obtidos utilizando-se as 18 sondas selecionadas. Um bom resultado também foi obtido para o modelo com 8 sondas.

Tabela 6.8: Resultados para o MOBJ-NN

Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.976	0.988	0.890	0.915	0.963	0.951
Se	0.905	0.952	0.714	0.714	0.857	0.810
Es	1.000	1.000	0.951	0.984	1.000	1.000
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.549	0.647	0.843	0.824	0.824	0.863
Se	0.769	0.615	0.538	0.615	0.692	0.769
Es	0.474	0.658	0.947	0.895	0.868	0.895

6.8 MOBJ-LASSO

O procedimento para o treinamento e para a escolha da solução foi o mesmo utilizado na seção 6.7. A utilização do algoritmo LASSO tem uma característica especial que é o *pruning* automático da rede quando o algoritmo passa por um vértice da restrição, fazendo que alguns pesos da rede sejam zero. Desta forma, observando os pesos da camada escondida consegue-se verificar quais atributos estão realmente sendo utilizados. Foi observado que em todos os experimentos a topologia resultante da rede consistiu em algumas entradas, apenas um neurônio na camada escondida e um neurônio na camada de saída. Assim tomou-se as médias dos pesos da camada escondida para as dez execuções do algoritmo e observou-se que aqueles pesos que tiveram um valor médio muito menor que os outros não foram utilizados com muita fre-

qüência pela rede, sendo assim desconsiderados. A Tabela 6.9 mostra, além dos resultados de classificação, o número médio de pesos que foram efetivamente utilizados pela rede.

Como pode-se observar, bons resultados ocorreram para os conjuntos de 8, 10, 32 e 18 sondas, sendo que o melhor resultado foi aquele obtido utilizando-se o conjunto de 18 sondas, já que para este conjunto o algoritmo utilizou somente 11 das sondas presentes.

Tabela 6.9: Resultados para o MOBJ-LASSO

Número médio de sondas efetivamente utilizados						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
No. S.	23	16	6	10	19	11
Conjunto de Treinamento						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	1.000	0.963	0.890	0.890	0.963	0.927
Se	1.000	0.905	0.714	0.714	0.857	0.809
Es	1.000	0.984	0.951	0.951	1.000	0.967
Conjunto de Teste						
	Natowicz	Hess	8 sondas	10 sondas	32 sondas	18 sondas
Ac	0.569	0.686	0.765	0.824	0.843	0.863
Se	0.769	0.615	0.462	0.615	0.769	0.769
Es	0.500	0.711	0.868	0.895	0.868	0.895

6.9 Avaliação dos Resultados

Para comparar todas as soluções simultaneamente foi calculada a distância ao ponto ótimo (0,1) do espaço ROC [2] tanto para o conjunto de treinamento quanto para o conjunto de teste, sendo os resultados apresentados na Figura 6.1. Essa forma de representar as soluções dá uma idéia da sensibilidade e da especificidade nos dois conjuntos [24]. A linha que conecta os pontos em que $d_{teste} = d_{treinamento}$ corresponde às soluções que tem a mesma performance tanto no conjunto de treinamento quanto de teste. Abaixo da linha encontram-se as soluções que tem performance melhor no conjunto de teste e acima as que tem performance melhor sobre o conjunto de treinamento. Desta forma, as melhores soluções são aquelas que estão mais próximas da origem do sistema de coordenadas e mais próximas da linha [24].

Como a Figura 6.1 contém muitos pontos a região onde se encontram as melhores soluções foi ampliada visando a facilitar a interpretação dos resultados. A ampliação dessa região encontra-se na Figura 6.2.

Como pode ser observado na Figura 6.2 as melhores soluções são: sondas de Natowicz com os classificadores de voto majoritário e DLDA, 32 sondas utilizando SVM com kernel RBF, 18 sondas utilizando Naïve Bayes, MLP e

SVM com kernel RBF e linear e 10 sondas utilizando SVM com kernel RBF e linear. A Tabela 6.10 mostra o número de falsos positivos e falsos negativos para cada um destes modelos e para o modelo de Hess et al. [19] usando o DLDA tanto para o conjunto de treinamento quanto para o conjunto de teste. Na Tabela também são apresentados os resultados para o modelo de 18 sondas com o classificador LASSO, já que este teve bons resultados e ainda reduziu o número de sondas necessárias para a classificação, como mostra a Tabela 6.9.

A Tabela 6.10 mostra que os resultados para o conjunto de teste foram muito próximos entre os modelos selecionados, diferindo em poucos pacientes. Já os resultados para o conjunto de treinamento são bem diferentes, apresentando soluções muito próximas da separação total e soluções com separação relativamente pior.

Observando a Figura 6.2 e os resultados apresentados na Tabela 6.10 pode-se concluir que o conjunto de sondas que apresenta resultados mais equilibrados, independente do classificador, é o conjunto de 18 sondas. Isso mostra que apesar do conjunto ter sido selecionado utilizando-se o classificador *naïve Bayes* ele funciona bem com outros classificadores diferentes.

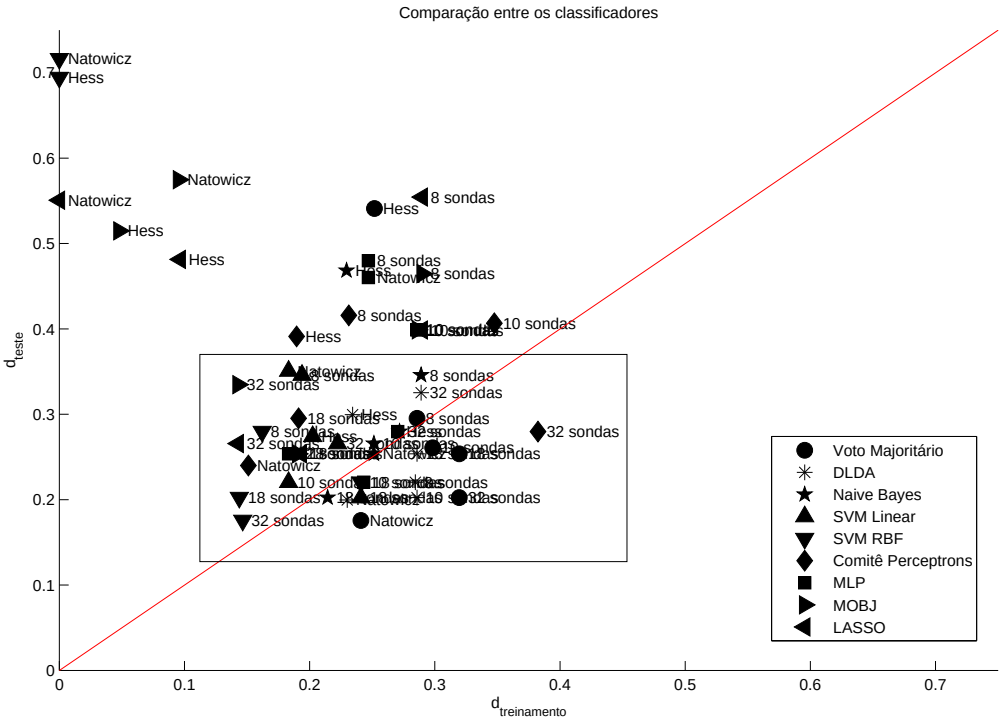


Figura 6.1: Distância no espaço ROC para os conjuntos de treinamento e teste

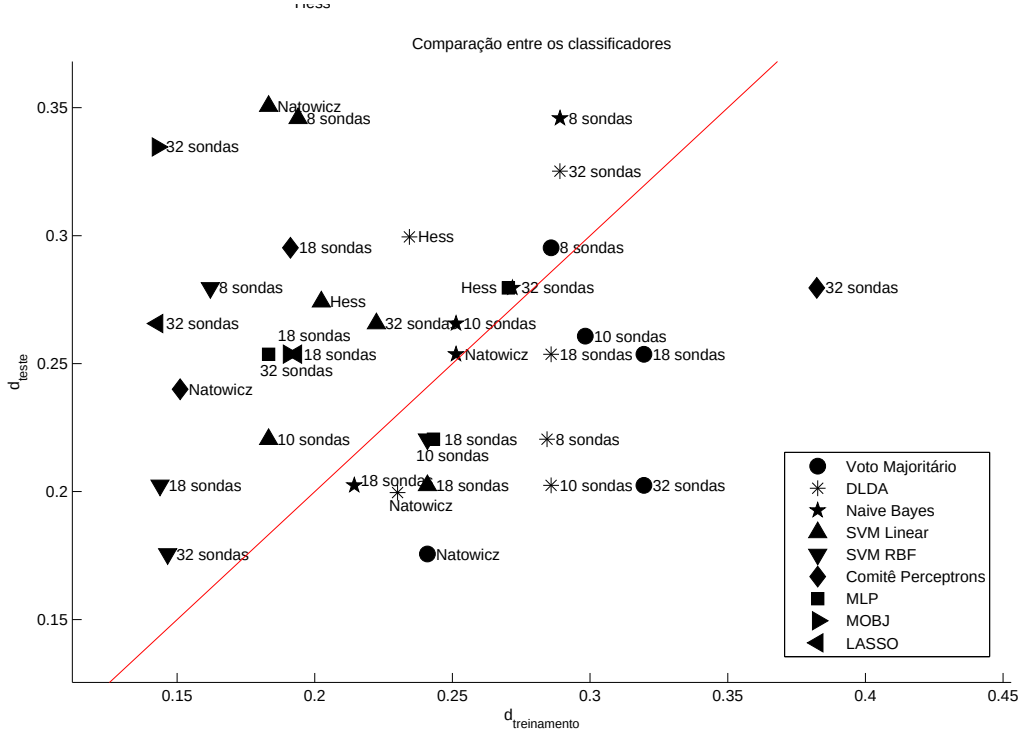


Figura 6.2: Zoom na região marcada pelo retângulo na Figura 6.1

Tabela 6.10: Número de falsos positivos e falsos negativos em cada base de dados

	Treinamento		Teste	
Modelo	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	1
Hess DLDA	14	1	11	1

Outro ponto interessante a ser analisado é o resultado obtido com as 18 sondas e o modelo LASSO, uma vez que este modelo reduziu o espaço de entrada de 18 para 11 sondas, obtendo bons resultados para o conjunto de treinamento e de teste. Este resultado pode levar à pergunta: porque não aplicar as 64 sondas selecionadas diretamente no modelo LASSO e deixar que o processo de otimização dos pesos selecione as sondas mais importantes no espaço de entrada da rede? A resposta a esta questão é bastante simples. Como o número de sondas selecionadas inicialmente é muito próximo

do número de padrões do conjunto de treinamento (82 pacientes) a solução escolhida pelo decisor de AUC apresentado nesta dissertação, tende a uma solução super-ajustada, pois a solução com maior AUC será aquela que separar corretamente todo o conjunto de treinamento. Já para um espaço de entrada reduzido como o de 18 sondas o modelo LASSO encontra bons resultados já que a chance da solução escolhida tender ao super-ajuste é menor, uma vez que o espaço de entrada é bem menor que o número de padrões de treinamento, tornando assim válida a seleção das 11 sondas mais significativas.

Como visto na seção 6.8, o número de entradas importantes das redes foram obtidos a partir das médias dos pesos da camada escondida ao final das 10 execuções do treinamento. Dessa forma foi possível estimar as entradas mais importantes, uma vez que a rede resultante tinha apenas um neurônio na camada escondida, o que resulta em apenas um peso para cada entrada. A Figura 6.3 mostra o valor médio de cada um dos pesos após as 10 execuções do algoritmo para o conjunto de 18 sondas e a Figura 6.4 mostra a topologia média da rede.

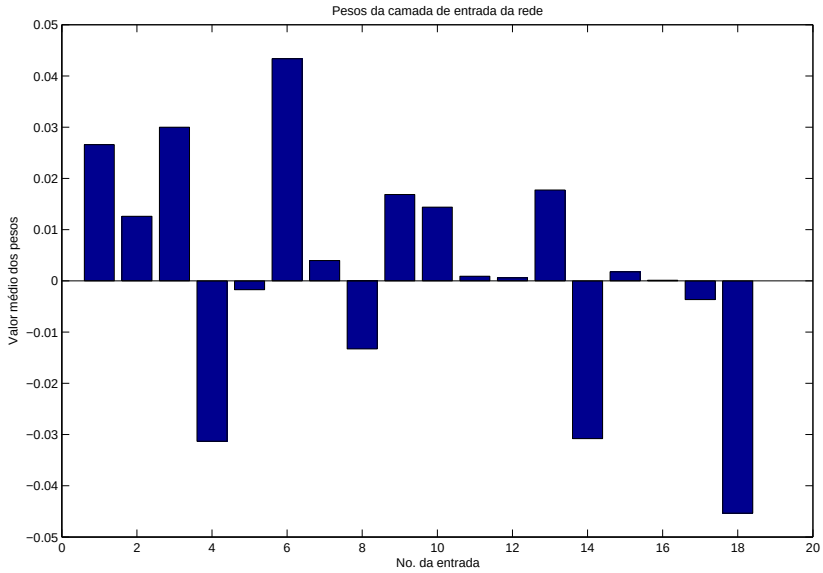


Figura 6.3: Gráfico de barras dos pesos médios da camada de entrada da rede

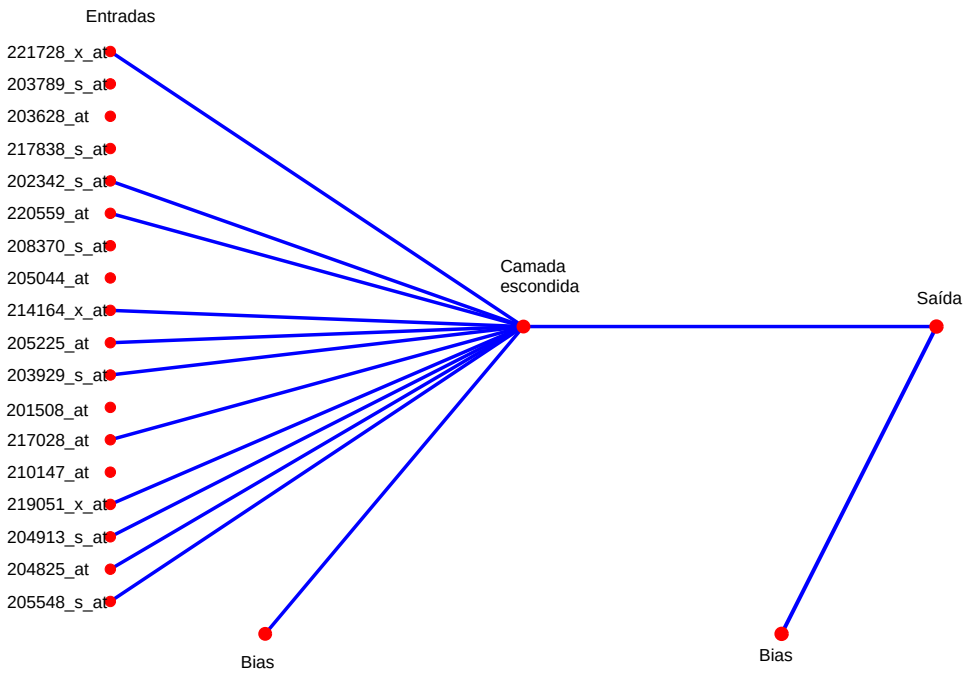


Figura 6.4: Topologia média da rede

Para verificar se as 11 sondas selecionadas pelo modelo LASSO funcionam bem para outros modelos, elas foram testadas em todos os classificadores utilizados nesta dissertação. Os resultados são apresentados na Tabela 6.11. Comparando a Tabela 6.12 com a Tabela 6.10 pode-se observar que o conjunto de 11 sondas obteve bons resultados, sendo que para o classificador *naïve Bayes* os resultados foram quase os mesmos que os obtidos com o conjunto de 18 sondas, diferindo em apenas 3 pacientes mal classificados no conjunto de treinamento. Esses resultados indicam que o conjunto de 11 sondas é suficiente para realizar a classificação dos pacientes. Esse conjunto tem a grande vantagem de ser 3 vezes menor que os conjuntos selecionados por Hess et al. [19] e por Natowicz et al. [27], o que reduz a chance de super-ajuste aos dados de treinamento e reduzindo também a chance das sondas serem obtidas ao mero acaso, aumentando a confiabilidade da solução.

Tabela 6.11: Resultados para as 11 sondas selecionadas pelo modelo LASSO

Conjunto de Treinamento									
	Voto	DLDA	Bayes	SVM L	SVM R	Per.	MLP	MOBJ	LASSO
Ac	0.793	0.805	0.841	0.915	0.854	0.890	0.951	0.951	0.927
Se	0.857	0.857	0.809	0.857	0.810	0.667	0.810	0.810	0.810
Es	0.770	0.787	0.852	0.934	0.869	0.967	1.000	1.000	0.967
Conjunto de Teste									
	Voto	DLDA	Bayes	SVM L	SVM R	Per.	MLP	MOBJ	LASSO
Ac	0.863	0.843	0.863	0.824	0.863	0.824	0.765	0.843	0.863
Se	0.769	0.769	0.846	0.923	0.846	0.615	0.538	0.7692	0.769
Es	0.895	0.868	0.868	0.789	0.868	0.895	0.842	0.8684	0.895

Tabela 6.12: Número de falsos positivos e falsos negativos em cada base de dados para o conjunto de 11 sondas

	Treinamento		Teste	
Modelo	FP	FN	FP	FN
Voto Majoritário	14	3	4	3
DLDA	13	3	5	3
Naïve Bayes	9	4	5	2
SVM Linear	4	3	8	1
SVM RBF	8	4	5	2
Comitê de Perceptrons	2	7	4	5
MLP	0	4	6	6
MOBJ	0	4	5	3
LASSO	2	4	4	3

6.10 Conclusões do capítulo

Neste capítulo foram apresentados os resultados de classificação para cada um dos conjuntos de sondas aplicados em cada um dos classificadores escolhidos para a dissertação. Foi visto que o modelo que obteve os melhores resultados foi o de 18 sondas aplicado ao classificador *naïve Bayes*. Foi verificado ainda que o modelo LASSO conseguiu reduzir o conjunto para 11 sondas e que este conjunto reduzido obteve excelentes resultados também para o classificador *naïve Bayes*, equivalentes àqueles obtidos pelo modelo de Natowicz et al. que utiliza 30 sondas. Observando esses dois resultados verificou-se que o modelo naïve Bayes com 18 sondas erra 3 pacientes a mais no conjunto de treinamento que o modelo LASSO com as 11 sondas, sendo que para o conjunto de teste os dois modelos variam em apenas um paciente em cada classe, mostrando que os dois resultados são equivalentes. Dessa forma concluí-se que uma boa classificação é possível utilizando apenas aproximadamente 1/3 do número de sondas apresentado por Natowicz et al., utilizando sondas na intersecção dos conjuntos e com novas sondas encontradas.

CAPÍTULO 7

Discussão

Como foi observado na seção anterior este trabalho indicou 2 conjuntos de sondas que parecem estar relacionados com o problema de previsão de resposta patológica completa (PCR). O primeiro conjunto contém 18 sondas e o segundo é um subconjunto do primeiro, contendo apenas 11 sondas. Foi observado ainda que um simples classificador *naïve Bayes* consegue bons resultados utilizando qualquer um destes dois conjuntos. Entretanto, é importante ressaltar que devido ao pequeno número de padrões para treinamento e para teste, algumas sondas diferentes podem entrar ou sair dos conjuntos se novos padrões forem acrescentados aos experimentos no futuro. Outra característica importante da técnica de seleção proposta nessa dissertação é a consideração de que os valores de cada sonda são normalmente distribuídos. A mesma consideração de normalidade foi feita por Natowicz et al. [27] no trabalho que deu origem ao projeto CAPES-COFECUB.

Uma análise importante é verificar quais sondas aparecem constantemente nas abordagens de seleção apresentadas, pois todas as técnicas utilizadas foram diferentes e certamente as sondas que aparecem na intersecção entre os conjuntos são as sondas mais informativas. Com este intuito a figura 7.1 mostra a intersecção dos conjuntos de sondas selecionadas por Hess et al. [19], Natowicz et al. [27] e as sondas selecionadas neste trabalho. Como pode ser observado, certamente as sondas mais informativas são 205548_s_at, 204825_at, 219051_x_at, já que elas aparecem em todos os métodos de seleção apresentados. As sondas 205548_s_at, 204825_at são as mais diferencialmente expressas na classe PCR (*down regulated*) e a sonda 219051_x_at é a mais diferencialmente expressa na classe NoPCR (*up regulated*) segundo o método *Volcano Plot* sendo que a primeira sonda é considerada a mais infor-

mativa segundo o valor de V(s) apresentado por Natowicz et al. [27]. A figura mostra ainda que o conjunto de 11 sondas selecionadas pelo método LASSO é constituído por 6 sondas que estão na intersecção dos conjuntos de Natowicz et al. [27] e de Hess et al. [19] com o conjunto de 18 sondas e por mais 5 sondas desse.

Na seção 6.9 foi visto que os resultados obtidos com o conjunto de 11 sondas são muito próximos daqueles obtidos por Natowicz et at. [27], mas com a grande vantagem de utilizar apenas um terço do número de sondas, sendo que daquele conjunto apenas 5 das 30 sondas selecionadas foram aproveitadas e as demais foram descobertas pelo método proposto nessa dissertação. Esse reduzido número de sondas facilita a análise das selecionadas em termos biológicos e reduz a chance do resultado ser uma mera combinação ao acaso de sondas que produzem bons resultados.

As Figuras 7.2(a) e 7.2(b) mostram dendogramas e *heat maps* gerados, respectivamente, sobre o conjunto de treinamento e de teste para as 18 sondas selecionadas. No *heat map* quanto mais forte for a cor vermelha mais expresso é a sonda e quanto mais forte for a cor verde menos expressa é a sonda. As setas vermelhas apontam para os pacientes que são falsos positivos e as setas verdes apontam para os pacientes que são falsos negativos, de acordo com o classificador *naïve Bayes*. Observando o padrão de nível de expressão apresentado nessas figuras fica claro o porquê desses pacientes terem tido uma classificação incorreta: eles têm o padrão de expressão invertido em quase todos os genes utilizados em relação ao padrão de expressão da classe verdadeira, ou seja, onde era esperada a cor vermelha na figura aparece a cor verde e onde era esperada a cor verde aparece a cor vermelha.

As Figuras 7.3(a) e 7.3(b) mostram as médias dos níveis de expressão para pacientes classificados corretamente (VP e VN) e classificados incorretamente (FN e FP). Comparando essas médias verifica-se novamente que o padrão dos níveis de expressão para pacientes classificados incorretamente estão invertidos em relação ao padrão daqueles classificados corretamente. Isso mostra a dificuldade do problema e a necessidade de se introduzir no futuro novas informações para tentar melhorar a classificação dos pacientes, como por exemplo informações clínicas. Neste trabalho não foi obtido êxito no uso das informações clínicas.

As Tabelas 7.1 e 7.2 mostram o nome de cada uma das 18 sondas e as características biológicas com as quais os genes representados pelas sondas já foram relacionados na literatura. Na tabela é dado destaque para as 11 sondas selecionadas pelo método LASSO que são mostradas em negrito.

Como pode ser observado, uma boa parte das sondas selecionadas estão relacionadas com processos ligados ao câncer [21], como apoptose (morte pro-

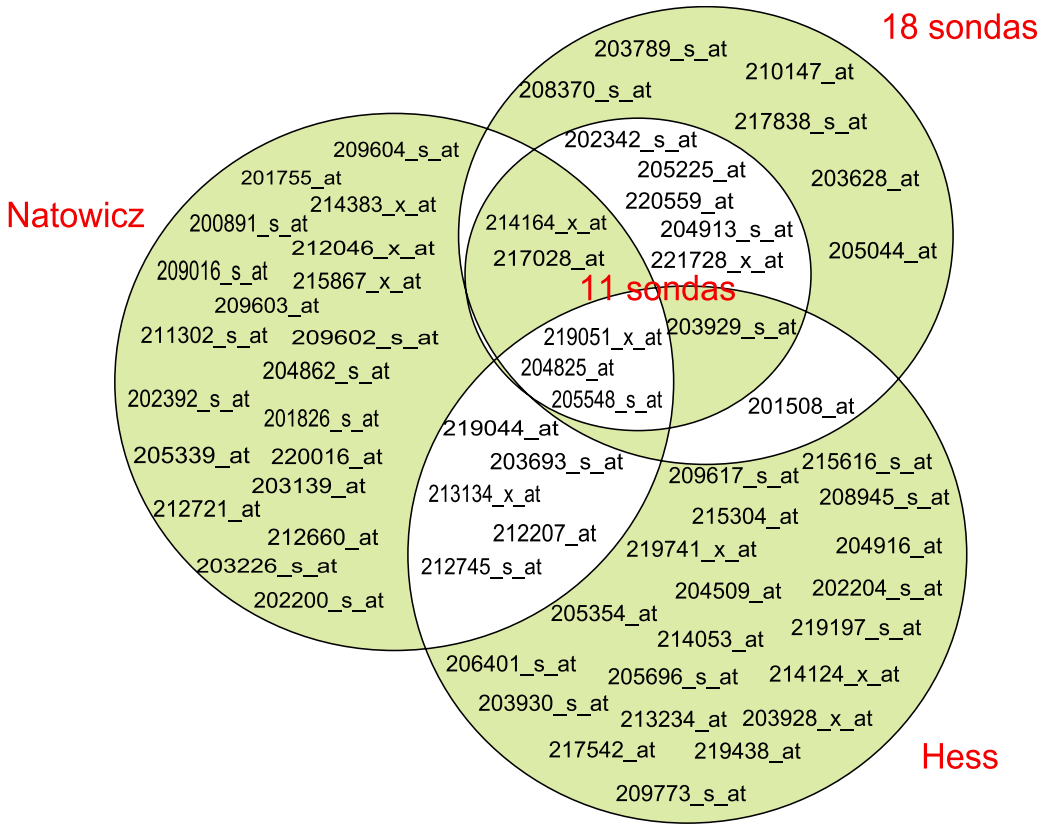
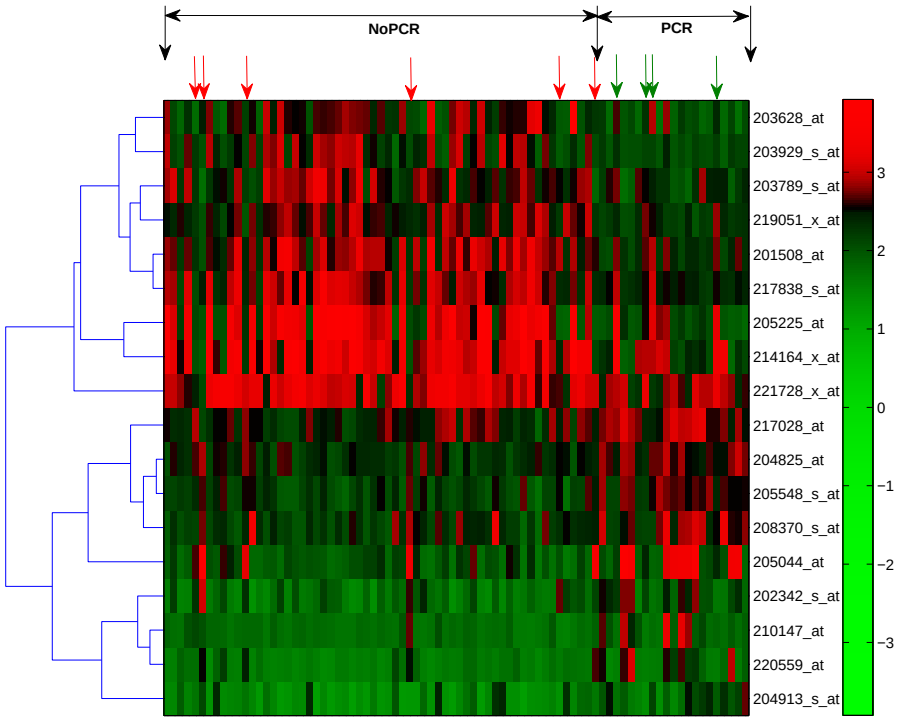


Figura 7.1: Intersecção entre os conjuntos de sondas selecionadas

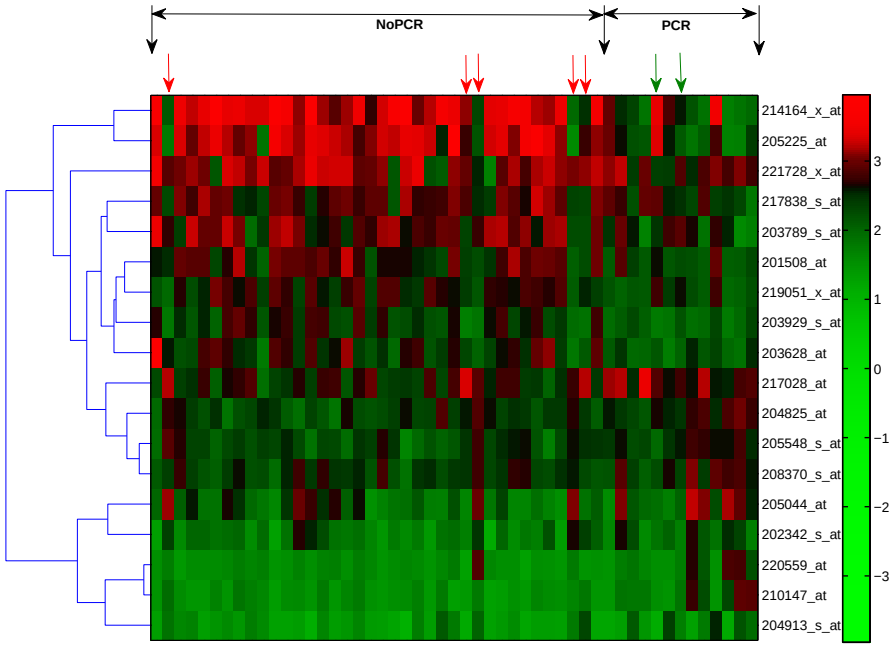
gramada das células), mitose (divisão celular), regulação do ciclo celular, transcrição e transdução do DNA e transdução de sinal, o que é um indicativo de que grande parte das sondas selecionadas são coerentes. Das 18 sondas selecionadas 13 tem características que podem estar relacionadas com o processo de desenvolvimento do câncer e das 11 destacadas 8 estão relacionadas.

7.1 Conclusões do capítulo

Nesse capítulo foram analisadas as 18 sondas selecionadas neste trabalho e foi verificado nas figuras de *heat map* que os pacientes que são classificados incorretamente possuem um padrão de nível de expressão invertido em relação ao padrão dos pacientes classificados corretamente. Foi visto ainda que certamente as sondas mais informativas são 205548_s_at, 204825_at, 219051_x_at, já que elas aparecem em todos os métodos de seleção apresentados como mostrado na figura 7.1. Por fim, a tabela 7.1 mostrou que, aparentemente, a maioria das sondas selecionadas estão relacionadas com processos biológicos ligados ao câncer.

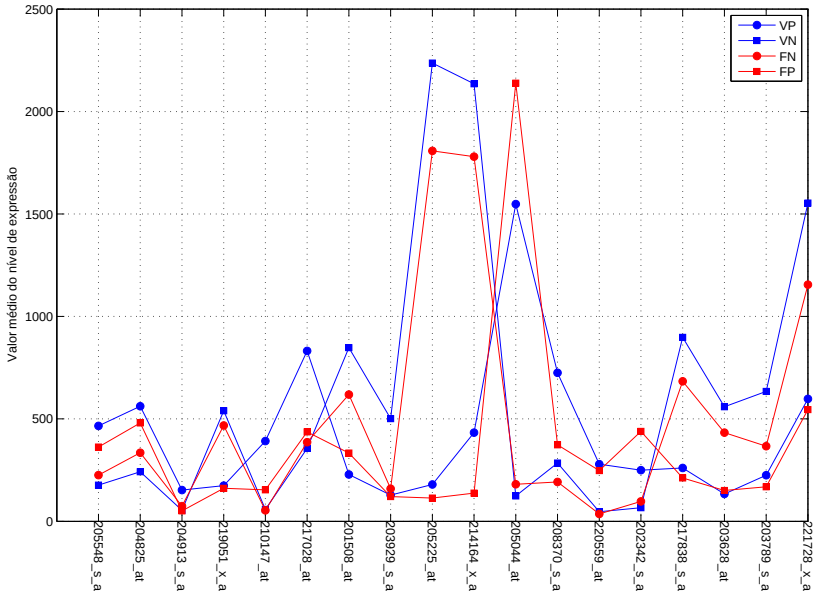


(a) *Heat map* e dendograma gerados sobre o conjunto de treinamento para as 18 sondas selecionadas

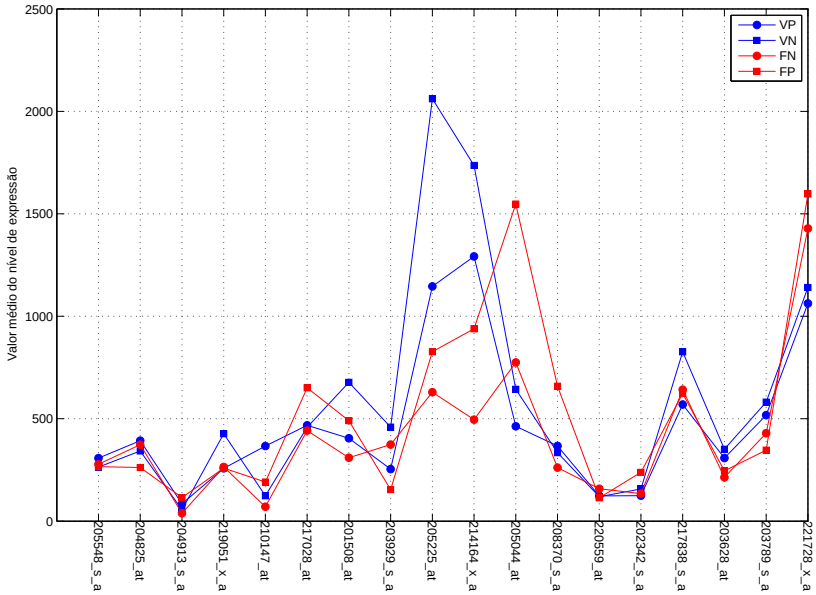


(b) *Heat map* e dendograma gerados sobre o conjunto de teste para as 18 sondas selecionadas

Figura 7.2: *Heat maps* para os conjuntos de treinamento e teste. As setas vermelhas apontam para os pacientes que são falsos positivos e as setas verdes apontam para os pacientes que são falsos negativos, de acordo com o classificador *naïve Bayes*



(a) Conjunto de treinamento



(b) Conjunto de teste

Figura 7.3: Comparação entre as médias dos níveis de expressão de cada sonda para pacientes VP, VN, FN e FP

Tabela 7.1: Sondas e genes correspondentes

Affymetrix Probe Set ID	**Mais expresso em**	**Gene Symbol**	**Gene Name**	**Relacionado com**
205548_s_at	PCR	BTG3	BTG family, member 3	regulação do ciclo celular, regulação negativa da proliferação celular
204825_at	PCR	MELK	maternal embryonic leucine zipper kinase	protein-tyrosine kinase activity, ATP binding, transferase activity
204913_s_at	PCR	SOX11	SRY (sex determining region Y)-box 11	regulação da transcrição
219051_x_at	NoPCR	METRN	meteorin	regulador da diferenciação de celulas *glial*
217028_at	PCR	CXCR4	chemokine (C-X-C motif) receptor 4	signal transduction, G-protein coupled receptor protein signaling pathway, integral to membrane
203929_s_at	NoPCR	MAPT	microtubule-associated protein τ	microtubule cytoskeleton organization and biogenesis, apoptosis
205225_at	NoPCR	ESR1	estrogen receptor 1	regulação da transcrição, crescimento celular, sinaliza o *pathway* de estrogênio receptor, regulação negativa da mitose
214164_x_at	NoPCR	CA12	carbonic anhydrase XII	one-carbon compound metabolism, integral to membrane
220559_at	PCR	EN1	engrailed homolog 1	regulação da transcrição, morphogenesis
202342_s_at	PCR	TRIM2	tripartite motif-containing 2	protein ubiquitination, ubiquitin ligase complex, cytoplasm
221728_x_at	NoPCR	XIST	X (inactive)-specific transcript	

Tabela 7.2: Sondas e genes correspondentes

Affymetrix Probe Set ID	**Mais expresso em**	**Gene Symbol**	**Gene Name**	**Relacionado com**
210147_at	PCR	ART3	ADP - ribosyl-transferase 3	protein amino acid ADP-ribosylation
201508_at	NoPCR	IGFBP4	insulin-like growth factor binding protein 4	regulação do crescimento celular
205044_at	PCR	GABRP	gamma-aminobutyric acid (GABA) A receptor, pi	ion transport, integral to membrane, postsynaptic membrane
208370_s_at	PCR	DSCR1	Down syndrome critical region gene 1	signal transduction, central nervous system development, circulation, calcium-mediated signaling
217838_s_at	NoPCR	EVL	Enah/Vasp-like	actin filament organization, cell surface receptor linked signal transduction, neurogenesis, axon guidance
203628_at	NoPCR	IGF1R	insulin-like growth factor 1 receptor	regulação do ciclo celular, anti-apoptose, regulação positiva da proliferação celular
203789_s_at	NoPCR	SEMA3C	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	immune response, transmembrane receptor protein tyrosine kinase signaling pathway, development, response to drug

CAPÍTULO 8

Conclusões e Trabalhos Futuros

8.1 Conclusões

Esta dissertação apresentou resultados para um problema de ordem social e estratégica para qualquer nação: o tratamento do câncer de mama. Nela buscou-se uma forma de prever se a quimioterapia pré-operatória teria boms resultados ou não para uma paciente, evitando assim que ela passasse por sofrimento desnecessário. Para tanto utilizou-se dados de nível de expressão gênica, buscando descobrir os genes que possuem um padrão de expressão para os casos em que a quimioterapia funciona e para os casos em que não funciona. A dissertação apresentou ainda o funcionamento do mecanismo geral de *microarrays* e o mecanismo específico de *microarray Affymetrix*.

Foram apresentados também os métodos de seleção de genes presentes na literatura que já foram aplicados aos dados da base utilizada nesta dissertação. Apresentou-se também uma grande variedade de classificadores com potencial para solucionar o problema de previsão de resposta patológica completa.

Mostrou-se ainda que é possível selecionar um conjunto de sondas significantes e capazes de classificar os pacientes apenas utilizando o conjunto de treinamento, o que evita que o resultado sofra *overfitting* nos dados de teste, já que na teoria de aprendizado de máquina o conjunto de teste não deve ser visto pelo classificador durante os seus ajustes. Observou-se que utilizando uma técnica de seleção amplamente utilizada na literatura [12] obtém-se resultados equivalentes a aqueles apresentados por Natowicz et al. [27], porém com um número de sondas quase 2 vezes menor. Verificamos ainda que o método LASSO consegue reduzir ainda mais o número de sondas, chegando a

quase 1/3 de sondas utilizadas por Natowicz et al. e por Hess et al.

Analisando os conjuntos de todas as sondas selecionadas pelos dois modelos da literatura e pelos dois modelos propostos na dissertação observou-se que as 3 sondas mais diferencialmente expressas no método *Volcano Plot* aparecem na intersecção entre todos os métodos, o que ressalta a importância das mesmas no problema de previsão de PCR.

Para conseguir resultados ainda melhores e mais confiáveis é necessário conseguir mais dados para serem utilizados na continuação do projeto CAPES-COFECUB.

8.2 Trabalhos Futuros

O maior problema enfrentado nessa pesquisa foi o reduzido número de padrões disponíveis na base de dados. Desta forma um dos principais passos do projeto CAPES-COFECUB será tentar conseguir recursos para coletar mais dados, de preferência no Brasil.

Com os dados disponíveis até o momento pode-se enumerar os seguintes passos:

1. Utilizar todas as sondas selecionadas pelo método *Volcano Plot* em um algoritmo genético para realizar a seleção das sondas mais importantes;
2. Aplicar outras técnicas estatísticas para a seleção das sondas mais significativas;
3. Aplicar outros classificadores nas sondas selecionadas;
4. Desenvolver novos classificadores;
5. Tentar integrar dados clínicos com dados genômicos;
6. Buscar uma solução eficiente em forma de árvore de decisão;
7. Buscar uma parceria com médicos da área de genômica, ginecologia ou oncologia.

Referências Bibliográficas

[1] Affymetrix. Affymetrix. http://www.affymetrix.com.

[2] W.S. Andrus and K.T. Bird. Radiology and receiver operating characteristic (roc) curve. *CHEST*, 67(4):378–379, 1975.

[3] M. Bazaras and C. M. Shetty. *Nonlinear Programming, Theory and Algorithms*. John Wiley & Sons, New York, 1979.

[4] J. C. Bennett and F. Cecil Plum and. *Tratado de Medicina Interna*. Guanabara Koogan, 1997.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] Robert G. Bland, Donald Goldfarb, and Michael J. Todd. The ellipsoid method: A survey. Technical report, Ithaca, NY, USA, 1980.

[7] 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)*, July 2008. Accepted.

[8] A.P. Braga, A.C.P.L.F. Carvalho, and T.B. Ludermir. *Redes Neurais Artificiais: teoria e aplicações*. Livros Técnicos e Científicos (LTC), Março 2000.

[9] Clinimater. http://www.clinimater.com.br.

[10] M. A. Costa and A. P. Braga. Optimization of neural networks with multiobjective lasso algorithm. In *IEEE World Congress on Computational Intelligence*, Vancouver, 2006.

[11] Marcelo Costa, Thiago Rodrigues, Euler Horta, Antônio Braga, Carmen D. M. Pataro, René Natowicz, Roberto Incitti, Roman Rouzier, and Arben Çela. *Learning and Approximation: Theoretical Foundations and Applications*, chapter New Multi-Objective Algorithms for Neural Network Training applied to Genomic Classification Data. Studies in Computational Intelligence. Springer Verlag, 2008.

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

[13] Instituto Nacional de Câncer. http://www.inca.gov.br.

[14] Alice de Medeiros Zelmanowicz. Câncer de mama. *ABC da Saúde*, 2001. http://www.abcdasaude.com.br.

[15] Antônio de Pádua Braga. Bayes classifiers. *Notas de aula da disciplina "Inteligência Computacional"*, 2007.

[16] Antônio de Pádua Braga, Ricardo H C Takahashi, Marcelo Azevedo Costa, and Roselito de Albuquerque Teixeira. *Multi-Objective Machine Learning*, chapter 7, pages 151–172. Studies in Computational Intelligence. Springer, 2006.

[17] F. Freitas, C. H. Menke, E. P. Passos, and W. A. Rivoire. *Rotinas Em Ginecologia*. ARTMED - BOOKMAN, 2005.

[18] 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.

[19] 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):4236–4244, 2006.

[20] Euler Guimarães Horta, Antônio de Pádua Braga, and Rodney Rezende Saldanha. Acelerando o treinamento multiobjetivo de rnas pelo método de gradiente projetado. *Congresso Brasileiro de Automática*, Setembro 2008.

[21] Lynn B. Jorde, John C. Carey, Michael J. Bamshad, and Raymond L. White. *Genética Médica*. Elsevier, 2004.

[22] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, December 1996.

[23] Olga Modlich, Hans-Bernd Prisack, Marc Munnes, Werner Audretsch, and Hans Bojar. Predictors of primary breast cancers responsiveness to preoperative epirubicin/cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. *Journal of Translational Medicine*, 3(1):32, 2005.

[24] R. Natowicz, A. P. Braga, R. Incitti, E. G. Horta, R. Rouzier, T. S. Rodrigues, and M. A. Costa. A new method of dna probes selection and its use with multi-objective neural networks for predicting the outcome of breast cancer preoperative chemotherapy. In *European Symposium on Neural Networks (ESANN08)*, 2008.

[25] René Natowicz, Roberto Incitti, Roman Rouzier, Arben Çela, Antônio Braga, Euler Horta, Thiago Rodrigues, and Marcelo Costa. *Computational Intelligence in human cancer research*, chapter Downsizing Multigenic Predictors of the Response to Preoperative Chemotherapy in Breast Cancer. Rapid Research Results. LNCS, 2008.

[26] René Natowicz, Roberto Incitti, Roman Rouzier, Arben Çela, Antônio Braga, Euler Horta, Thiago Rodrigues, and Marcelo Costa. Downsizing multigenic predictors of the response to preoperative chemotherapy in breast cancer. *In 12th International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES08)*, September 2008.

[27] 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.

[28] Medline Plus. Breast cancer tutorial. http://www.nlm.nih.gov/medlineplus/tutorials/breastcancer/htm/index.htm.

[29] J. B. Rosen. The gradient projection method for nonlinear programming. part I. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):181–217, mar 1960.

[30] J. B. Rosen. The gradient projection method for nonlinear programming. part II. nonlinear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):514–532, dec 1961.

[31] RR Saldanha, RHC Takahashi, JA Vasconcelos, and JA Ramirez. Adaptive deep-cut method in ellipsoidal optimization for electromagnetic design. *IEEE Transactions on Magnetics*, 35(3):1746–1749, may 1999.

[32] E. Schadt, C. Li, B. Ellis, and W. Wong. Feature extraction and normalization algorithm for high-density oligonucleotide gene expression array data, 2001.

[33] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, October 1995.

[34] Robinson Semolini. Support vector machines, inferência transdutiva e o problema de classificação, Dezembro 2002.

[35] 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.

[36] American Cancer Society. http://www.cancer.org.

[37] Dov Stekel. *Microarray Bioinformatics*. Cambridge University Press, 2003.

[38] R. H. C. Takahashi. *Otimização Escalar e Vetorial: Notas de aula*. Belo Horizonte, 2004.

[39] R. A. Teixeira. *Treinamento de Redes Neurais Artificiais Através de Otimização Multi-Objetivo: Uma Nova Abordagem para o Equilíbrio entre a Polarização e a Variância*. PhD thesis, CPDEE, UFMG, 2001.

[40] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

[41] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.