Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Curso de Engenharia Elétrica

Classificação de Lesões em Mamografias por Análise de Componentes Independentes, Análise Discriminante Linear e Máquina de Vetor de Suporte

---

Daniel Duarte Costa

São Luís
2008

Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Curso de Engenharia Elétrica

Classificação de Lesões em Mamografias por Análise de Componentes Independentes, Análise Discriminante Linear e Máquina de Vetor de Suporte

---

Daniel Duarte Costa

Dissertação apresentada ao Curso de
Pós-Graduação em Engenharia de Eletricidade da UFMA
como parte dos requisitos necessários para obtenção do
grau de Mestre em Engenharia Elétrica.

São Luís
2008

Costa, Daniel Duarte

Classificação de Lesões em Mamografias por Análise de Componentes Independentes, Análise Discriminante Linear e Máquinas de Vetor de Suporte. / Daniel Duarte Costa. - São Luís, 2008.

56f.:il.

Dissertação (Mestrado em Engenharia de Eletricidade) - Centro de Ciências Exatas e Tecnologia, Universidade Federal do Maranhão, 2008.

1.Processamento de imagens médicas. 2.Diagnóstico auxiliado por computador. 3.Mamografias - análise de imagens.. I.Barros, Allan Kardec, orient. II. Título.

CDU 004.932:61

Classificação de Lesões em Mamografias por Análise de Componentes Independentes, Análise Discriminante Linear e Máquina de Vetor de Suporte

Daniel Duarte Costa

Aprovado em 25/02/2008

BANCA EXAMINADORA

Prof. Allan Kardec Duailibe Barros Filho

Dr. em Engenharia Elétrica - UFMA

Orientador

Prof. Hani Camille Yehia

Dr. em Engenharia - UFMG

Examinador Externo

Prof. João Viana da Fonseca Neto

Dr. em Engenharia Elétrica - UFMA

Examinador Interno

A Deus, fonte da vida.
Aos meus pais Lindemberg e Noêmia, pelo incentivo e carinho constantes.
A Ana Cláudia, minha noiva, pela compreensão e pela paciência.
Aos amigos, pelo apoio e companheirismo.
Aos professores e funcionários do Departamento de Engenharia Elétrica.

Agradecimentos

À força criadora da vida;

À minha família;

À minha noiva, Ana Cláudia, sempre compreensiva e amorosa.

Aos meus amigos do Laboratório de Processamento da Informação Biológica (PIB): André Cavalcante, Enio Aguiar, Fábio Marques, Cristiane Cristina, Flávio Mello, áurea Celeste, Sidcley, Márcio, Deusdete Brito, Ewaldo Santana, Eder Júnior, Aline, Isabela Bispo, Anderson, Euler Nicolal, Lucas Valadão, Diego, entre outros.

A um grande amigo, Geraldo Braz Júnior, por estar sempre me ajudando em vários momentos.

Aos professores Dr. Allan Kardec Barros e Dr. Aristófanes Corrêa Silva, pelo apoio, paciência, competência e dedicação.

A todos que, direta ou indireta, contribuíram para a elaboração deste trabalho.

"Só atingiria sucesso na vida pela autodisciplina. Apliquei-a até que meus desejos se realizassem."

Nicola Tesla

Resumo

Câncer de mama feminino é o câncer que mais causa morte nos países ocidentais. Esforços em processamento de imagens foram feitos para melhorar a precisão dos diagnósticos por radiologistas. Neste trabalho, nós apresentamos uma metodologia que usa análise de componentes independentes (ICA) junto com análise discriminante linear (LDA) e máquina de vetor de suporte (SVM) para distinguir as imagens entre nódulos ou não-nódulos e os tecidos em benignos ou malignos. Como resultado, obteve-se com LDA 90,11% de acurácia na discriminação entre nódulo ou não-nódulo e 95,38% na discriminação de tecidos benignos ou malignos na base de dados DDSM. Na base de dados mini-MIAS, obteve-se 85% e 92% na discriminação entre nódulos ou não-nódulos e tecidos benignos ou malignos respectivamente. Com SVM, alcançou-se uma taxa de até 99,55% na discriminação de nódulos ou não-nódulos e a mesma porcentagem na discriminação entre tecidos benignos ou malignos na base de dados DDSM enquanto que na base de dados mini-MIAS, obteve-se 98% e até 100% na discriminação de nódulos ou não-nódulos e tecidos benignos ou malignos, respectivamente.

Palavras-chave: análise de componentes principais, análise de componentes independentes, análise discriminante linear, diagnóstico auxiliado por computador, mamografias, máquinas de vetor de suporte.

Abstract

Female breast cancer is the major cause of death in western countries. Efforts in Computer Vision have been made in order to add improve the diagnostic accuracy by radiologists. In this work, we present a methodology that uses independent component analysis (ICA) along with support vector machine (SVM) and linear discriminant analysis (LDA) to distinguish between mass or non-mass and benign or malign tissues from mammograms. As a result, it was found that: LDA reaches 90,11% of accuracy to discriminante between mass or non-mass and 95,38% to discriminate between benign or malignant tissues in DDSM database and in mini-MIAS database we obtained 85% to discriminate between mass or non-mass and 92% of accuracy to discriminate between benign or malignant tissues; SVM reaches 99,55% of accuracy to discriminate between mass or non-mass and the same percentage to discriminate between benign or malignat tissues in DDSM database whereas, and in MIAS database it was obtained 98% to discriminate between mass or non-mass and 100% to discriminate between benign or malignant tissues.

Word-key: principal component analysis, independent component analysis, linear discriminant analysis, computer aided diagnosis, mammogram, support vector machine.

Lista de Tabelas

Lista de Figuras

Lista de Siglas

[BSS] - *Blind source separation* (Separação cega de fontes)
[CAD] - *Computer-aided diagnosis* (Diagnóstico auxiliado por computador)
[DDSM] - *Database for screening mammography*
[ICA] - *Independent components analysis* (Análise de componentes independentes)
[INCA] - Instituto nacional do câncer
[LDA] - *Linear discriminant analysis* (Análise discriminante linear)
[MIAS] - *Mammographic image analysis society*
[PCA] - *Principal components analysis* (Análise de componentes principais)
[ROI] - *Region of interest* (Região de interesse)
[SVM] - *Support vector machine* (Máquina de vetor de suporte)

Sumário

Capítulo 1

Introdução

O câncer de mama é o segundo tipo de câncer mais freqüente no mundo, perdendo apenas para o câncer de pele do tipo não melanona, e o mais comum entre as mulheres. A cada ano, cerca de 22% dos casos novos de câncer em mulheres são de mama. No Brasil, no ano de 2008, são esperados 49.400 novos casos, com um risco estimado de 51 casos a cada 100 mil mulheres. No estado do Maranhão, o risco estimado é de 310 novos casos a cada 100 mil mulheres [INCA, 2008].

De acordo com o Instituto Nacional do Câncer Americano é estimado que a cada três minutos uma mulher é diagnosticada com câncer de mama e a cada 13 minutos uma mulher morre devido a esta doença [NCI, 2006].

Além do fator sexo feminino, a idade é o fator de risco mais importante no câncer de mama. O risco também aumenta com mutações genéticas hereditárias nos genes BRCA1 e BRCA2, histórico pessoal ou familiar de câncer de mama, alta densidade no tecido mamário (uma medida mamográfica da quantidade de tecido glandular em relação ao tecido gorduroso da mama), biópsia confirmada de hiperplasia (especialmente hiperplasia atípica) e altas doses de radiação no peito como resultado de procedimentos médicos [Thuler, 2003].

Os fatores de risco relacionados à vida reprodutiva da mulher, como a menarca precoce, nuliparidade (maior número de ovulações), primeira gestação tardia (acima dos 30 anos), anticoncepcionais orais, menopausa tardia e terapia de reposição hormonal, estão bem estabelecidos em relação ao desenvolvimento do câncer de mama [Thuler, 2003].

O câncer de mama, assim como os demais, é resultado de alterações do DNA, que levam a uma proliferação celular desordenada. Quando há uma falha do mecanismo regulador que mantém o equilíbrio entre o crescimento celular e o bem estar do organismo, ocorre a formação de massas, denominadas tumores ou neoplasias, as quais são classificadas como malignas ou benignas. As neoplasias benignas têm crescimento organizado, em geral lento, e o tumor apresenta contorno bem nítido. Na neoplasia maligna, o crescimento é rápido, desordenado e infiltrativo. Nos tumores malignos, suas células têm capacidade de se desenvolver em outras partes do corpo, fenômeno este denominado metástase [Bauer et al.,1980].

A prevenção primária dessa neoplasia ainda não é totalmente possível devido à variação dos fatores de risco e às características genéticas que estão envolvidas na sua etiologia. Novas estratégias de rastreamento factíveis para países com dificuldades orçamentárias têm sido estudadas, uma vez que até o momento é indicada a mamografia para mulheres com idade entre 50 e 69 anos como método efetivo para detecção precoce.

No Brasil, o ministério da saúde recomenda como principais estratégias de rastreamento populacional um exame mamográfico a cada dois anos, para mulheres de 50 a 69 anos de idade, e o exame clínico anual das mamas, para mulheres de 40 a 49 anos de idade. O exame clínico da mama deve ser realizado em todas as mulheres que procuram o serviço de saúde, independentemente da faixa etária, como parte do atendimento à saúde da mulher. Para mulheres de grupos populacionais considerados de risco elevado para o câncer de mama (com histórico familiar de câncer de mama em parentes de primeiro grau), recomenda-se o exame clínico da mama e a mamografia, anualmente, a partir de 35 anos de idade.

A sobrevida das pacientes é diretamente relacionada com o tamanho do tumor no diagnóstico inicial. O diagnóstico precoce não apenas influencia o prognóstico, mas propicia cirurgia menos mutilante e com sobrevida comparável a intervenções cirúrgicas mais dramáticas e agressivas [Bauer et al.,1980].

Apesar de ser considerado como um câncer de relativamente bom prognóstico, se diagnosticado e tratado oportunamente, as taxas de mortalidade por câncer de mama continuam elevadas no Brasil, muito provavelmente porque a doença ainda

seja diagnosticada em estágios avançados. Nesses estágios, a doença já evoluiu de forma que o organismo não é capaz de responder ao tratamento, que em geral é mutilante e causa maior sofrimento à mulher. Com base nos dados disponíveis de Registros Hospitalares, 60% dos tumores de mama, em média, são diagnosticados em estágios III e IV [INCA, 2008].

1.1 Mamografia e o câncer de mama

O câncer de mama é a maior causa de mortes de câncer na população feminina. E já se sabe que o melhor método de prevenção é o diagnóstico precoce, que diminui a mortalidade e aumenta a eficácia do tratamento [INCA, 2008]. Portanto, um grande esforço tem sido feito para melhorar as técnicas de diagnóstico precoce. Entre elas, a mais utilizada é a mamografia, que é um método simples, barato e acessível.

O primeiro sinal de um câncer de mama é normalmente uma anormalidade detectada na mamamografia antes que possa ser sentida pela própria mulher ou por um agente de saúde. Grandes tumores podem ser evidenciados como um nódulo indolor. Sintomas como mudanças persistentes na mama, como espessamento, edema, distorção, sensibilidade, irritação da pele, *scaliness* ou anormalidades no bico como uma ulceração, retração ou descargas espontâneas são menos comuns. Tipicamente, dores mamárias vêm de condições benignas e isto não é um sintoma inicial do câncer de mama [ACS, 2007].

Pelo exame mamográfico pode-se detectar o câncer de mama em um estágio inicial quando o tratamento pode ser mais eficiente. Vários estudos têm demonstrado que a detecção precoce salva vidas e aumenta as opções de tratamento. As recentes quedas nas mortalidades por câncer de mama em mulheres são atribuídas à combinação da detecção precoce e do melhoramento do tratamento. A mamografia tem um alto nível de exatidão mas como podemos observar através dos testes médicos, ela não é perfeita. Em média, a mamografia detecta cerca de 80% a 90% dos cânceres de mama em mulheres sem sintomas [ACS, 2007].

A mamografia constitui uma forma particular de radiografia, que trabalha com níveis de tensões e correntes em intervalos específicos, destinada a registrar

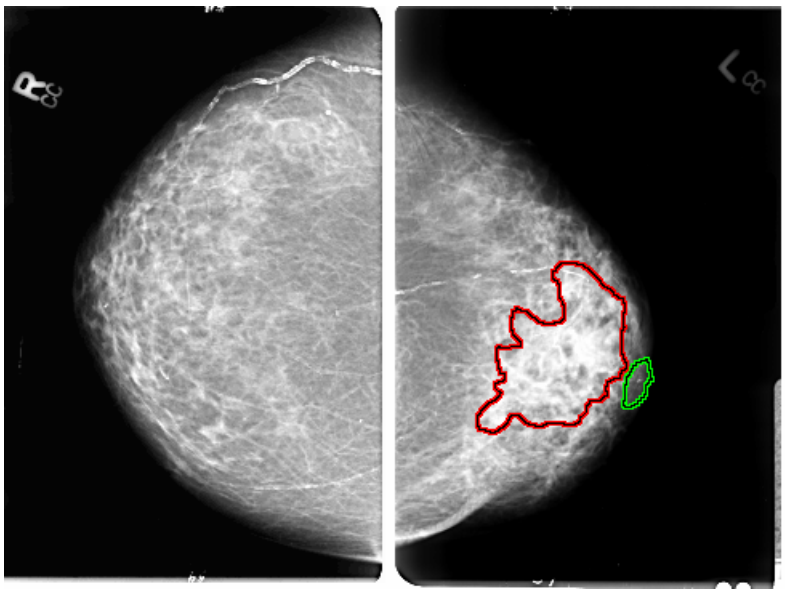

Figura 1.1: Incidência médio-lateral oblíqua com dois tumores malignos devidamente marcados na mama esquerda de uma paciente de 65 anos. Fonte: banco de dados DDSM [Heath et al., 1998]. Indentificação Volume: câncer_01 Caso: B-3027-1.

imagens da mama a fim de diagnosticar a presença ou ausência de estruturas que possam indicar patologias. Segundo o INCA, em uma mamografia, duas incidências de cada mama são indispensáveis: uma visão lateral ou oblíqua, como mostra a Figura 1.1 e uma crânio-caudal, conforme a Figura 1.2.

No entanto, a incidência médio-lateral-oblíqua (MLO) é a mais eficaz, pois mostra uma quantidade maior de tecido mamário e inclui estruturas mais profundas do quadrante superior externo e do prolongamento axilar.

A incidência crânio-caudal (CC) tem como objetivo incluir todo o material póstero-medial, completando a médio-lateral-oblíqua, que com freqüência não está totalmente demonstrado na incidência MLO. Permite também mais compressão da mama, uma vez que não inclui a axila, resultando em uma definição superior da arquitetura mamária e de lesões. Os radiologistas estudam as incidências crânio-caudais e as médio-laterais aos pares de modo a permitir a comparação de regiões simétricas, pois qualquer assimetria pode ser indício de patologia.

A mamografia é um exame de alta sensibilidade, apesar de a maioria dos estudos evidenciarem perdas entre 10% a 15% [INCA, 2008] dos casos de câncer

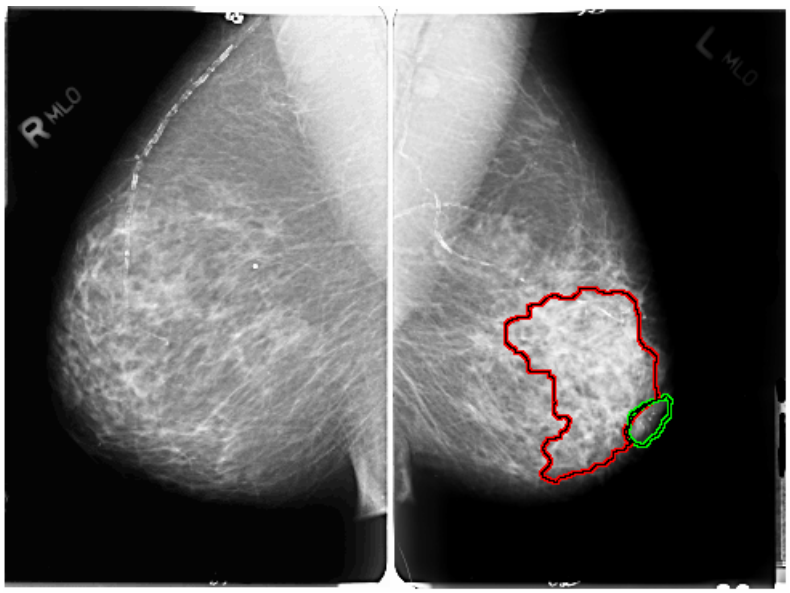

Figura 1.2: Incidência crânio-caudal com dois tumores malignos devidamente marcados na mama esquerda de uma paciente de 65 anos. Fonte: banco de dados DDSM [Heath et al., 1998]. Indentificação Volume: câncer_01 Caso: B-3027-1.

com tumor detectável ao exame clínico. Esta sensibilidade, no entanto, está diretamente relacionada à idade da mulher, sendo muito menor nas mulheres jovens, que apresentam uma alta densidade de tecido mamário, devido à predominância de tecidos fibroglandulares na sua composição.

O reconhecimento de estruturas que possam indicar a presença de câncer ocorre através da constatação de uma diferença de contraste entre os diversos tecidos envolvidos. A gordura, por exemplo, absorve uma menor quantidade de raios-X, aparecendo mais escura no mamograma, enquanto tecidos fibroglandulares apresentam densidade óptica maior e aparecem mais claros [Boyd et al.,1995]. Geralmente, microcalcificações e massas aparecem em tonalidades mais claras na imagem obtida após a revelação do filme mamográfico, mas essa diferenciação fica prejudicada em imagens de mamas densas. Por esse motivo, muitas vezes a descoberta do câncer de mama em mulheres com menos de 40 anos de idade acontece quando o tumor já apresenta um desenvolvimento avançando, o que dificulta o tratamento da doença. O diagnóstico de carcinomas não-palpáveis só é possível através da realização de mamografias minuciosas, em que cada detalhe é de extrema importância para evitar os diagnósticos falso-positivos e

falso-negativos [Bland e Copeland,2000].

1.2 Diagnóstico auxiliado por computador

Sistemas de diagnóstico auxiliado por computador (CAD, de "computer-aided diagnosis") vêm sendo desenvolvidos por diversos grupos de pesquisa, visando auxiliar a detecção precoce do câncer de mama para que a paciente possa ter o tratamento adequando o mais cedo possível. Pois é sabido que a descoberta da doença na fase inicial favorece a sua cura.

O objetivo principal destes sistemas é fornecer aos médicos uma segunda opinião ou uma sugestão de diagnóstico. Os CADs contornam problemas que surgem da subjetividade de um laudo humano, tais como expectativas, pré-conceitos e cansaço que interferem no diagnóstico do radiologista. Buscam assim, abaixar as taxas de falsos positivos, ou seja, casos em que a suspeita de malignidade leva a mulher à biópsia - exame altamente invasivo e de maior custo - para um resultado negativo. Vários estudos revelam que algoritmos automáticos de detecção são capazes de proporcionar um aumento de mais de 20% no total de acertos do radiologista e, com isso, consegue-se evitar biópsias desnecessárias [Zhang et al., 2002] [Feig, 2006].

Para que isto ocorra, é importante desenvolver técnicas para detectar e reconhecer lesões e regiões suspeitas e discriminá-las. Vários métodos de diagnósticos de patologias em mamogramas têm sido desenvolvidos por diferentes grupos de pesquisadores, contribuindo para um diagnóstico mais confiável.

Em [Zhang et al., 2004] foi proposto um algoritmo de rede neural que realiza em conjunto a seleção de característica e a classificação. A metodologia alcançou 87,2% de acerto para classificação de massas em diferentes subconjuntos de testes. Outro trabalho com objetivo semelhante é [Moayedi et al., 2007] onde o autor utilizou SVM (máquinas de vetor de suporte) baseado em uma rede neuro-fuzzy para desempenhar as tarefas do classificador, com características extraídas a partir da representação das massas no domínio da freqüência utilizando os coeficientes obtidos a partir de *contourlet* [Minh N. Do e Martin Vetterli, 2005]. Uma nova abordagem em redes neurais foi apresentada em [Lim e Er, 2004] com uma rede neuro-fuzzy genérica e dinâmica (GDFNN) para classificação de massas

em mamografias descritas por matriz de co-ocorrencia e distribuição do gradiente, alcançando acurácia de 70%.

Com objetivo semelhante, a metodologia proposta em [Martins et al., 2006] atingiu 86,85% de acerto para classificação de ROIs (região de interesse) usando uma rede neural Bayesiana. Em [Oliver et al. 2006] foi apresentada uma nova abordagem baseada no cálculo de autovalores e autovetores, feito tipicamente em sitemas de detecção de faces para classificação, de massas e tecidos normais em mamografias.

Em [Chen e Chang, 2004] foi proposta uma nova aplicação de um extrator de textura TUC (*Texture Unit Coding*). [Timp et al., 2007] analisou o desempenho da metodologia de classificação de masssas quando as características extraídas são realizadas em mamgorafias consecutivamente obtidas no tempo. O principal objetivo foi melhorar a descrição de massas utilizando informações presentes em mais de uma mamografia obtidas sucessivamente.

Em [Wei et al., 1995] foi investigada a capacidade de análise em multiresolução a partir da transformada *wavelet* para classificação de regiões de mamografias em massa e normal. Os resultados obtidos comprovam que os coeficientes da *wavelet* condensam com eficiência as características do tecido presente na região. Em [Masotti, 2006] foi desenvolvida uma nova abordagem não-paramétrica, de orientação seletiva, e em multiresolução nomeada representação de imagem *ranklet*. Em [Braz et al., 2007b] também foi verificado que a análise em multiresolução melhora o desempenho da extração de características para metodologias de classificação de massas em mamografias obtendo acerto total igual a 98,36%, superior aos resultados obtidos em [Braz et al., 2007a] para classificação de massas e não-massas sem utilizar a análise em multiresolução.

No trabalho [Campanini et al, 2002], mamografias foram classificadas em malignas e benignas através de SVM. Neste trabalho, os autores aplicaram a transformada de *wavelet*, e a partir de cada característica extraída foi utilizado SVM para a classificação final. O sistema obteve uma sensibilidade de 84%.

Em [Christoyianni et al., 2002] comparou-se três métodos: *Gray level histogram moments* (GLHM), que utiliza medidas estatísticas como média, desvio padrão, variância, assimetria, curtose, etc, as quais servem como parâmetros de entrada para um classificador baseado em redes neurais. *Spacial gray level*

dependence matrix (SGLD) que é construída a partir da contagem do número pares de pixels em uma dada janela. E análise de componentes independentes (ICA), em que o autor utiliza previamente análise de componentes principais (PCA) para reduzir o número de componentes e, logo após, utiliza ICA para extrair características significativas de cada imagem, para posteriomente classificá-las, com uma rede neural artificial. De acordo com os autores, ICA obteve a melhor performance, com 88% de sucesso discriminando os mamogramas entre normais e anormais e 79,31% discriminando entre normal, benigno e maligno.

Em [Campos et al., 2005a] e [Campos et al., 2005b] foi utilizada ICA para extração de características e uma rede neural multicamadas perceptron para classificação de regiões da mamografia. A metodologia alcançou 97,83% de acerto. Em [Costa et al., 2006] foi comparada a eficiência dos classificadores de SVM e LDA (Análise Discriminante Linear), após a extração de características utilizando ICA. Foi obtido até 99,6% de acerto na classificação.

Os trabalhos relacionados acima motivam a realização de uma investigação de análise de componentes independentes como extrator de características e a LDA e o SVM como classificadores, já que em outros trabalhos eles obtiveram desempenhos satisfatórios.

1.3 Objetivo

O objetivo deste trabalho é propor um método de classificação de regiões de interesses (ROI) de mamografias em normais ou anormais e em benignos ou malignos, para auxiliar radiologistas a determinarem possíveis patologias no exame clínico das mamas. Para alcançarmos nosso objetivo utilizamos análise de componentes principais (PCA, de *principal components analysis*) para redução de dimensionalidade e análise de componentes independentes (ICA, de *independent component analysis*) para a extração de características. Em seguida estas características serão usadas como parâmetros de entrada para um classificador que efetuará a classificação final.

1.4 Organização do trabalho

Este trabalho está organizado da seguinte forma:

No Capítulo 2, são descritos alguns conceitos e fundamentos teóricos utilizados neste trabalho, tais como: processamento de imagens, análise de componentes principais, análise de componentes independentes, análise discriminante linear e máquinas de vetor de suporte.

Os materiais e métodos utilizados neste trabalho serão mostrados no Capítulo 3.

No Capítulo 4 mostraremos os resultados da aplicação da metodologia e faremos algumas discussões.

Para finalizar, o Capítulo 5 constará das conclusões e das propostas de trabalhos futuros.

Capítulo 2

Fundamentos teóricos

2.1 Processamento digital de imagens

O termo processamento digital de imagens geralmente se refere ao processamento de uma imagem por um computador. Isto implica em um processamento digital de dados bidimensionais. Um imagem digital é um vetor de números reais ou complexos representados por um número finito de bits [Jain, 1989].

O processamento de imagens digitais abrange uma ampla escala de hardware, software e fundamentos teóricos [Gonzales e Woods, 2001]. Vamos discutir os passos fundamentais para executar uma tarefa de processamento de imagens.

O primeiro passo no processo é a aquisição da imagem, isto é, adquirir uma imagem digital. Para fazer isso necessitamos de um sensor para imageamento e a capacidade de digitalizar o sinal produzido pelo sensor. Este sensor pode ser uma câmera ou um scanner. O tipo de sensor a ser utilizado varia conforme a aplicação. Em nosso caso, foram utilizados scanners de alta precisão para digitalizar as imagens mamográficas.

O passo seguinte é o pré-processamento. Esta é a fase de melhoramento da imagem para que as chances de sucesso dos processos seguintes sejam maiores. Como exemplo de pré-processamento, temos o realce de contraste e a redução de dimensionalidade, ambos usados neste trabalho.

O terceiro passo é o da segmentação. Definida em termos gerais, a segmentação divide uma imagem de entrada em partes ou objetos constituintes. Em geral,

a segmentação automática é uma das tarefas mais difíceis no processamento de imagens digitais. Por um lado, um procedimento de segmentação robusto favorece substancialmente a solução bem sucedida de um problema de imageamento. Por outro lado, algoritmos de segmentação fracos ou erráticos quase sempre asseveram falha no processamento. No caso da classificação de nódulos mamários, a segmentação deve extrair os nódulos, isto é, a região de interesse, e descartar todo o resto da mamografia. Neste trabalho, a segmentação foi realizada de forma manual.

O processo de descrição, também chamado seleção de característica, procura extrair características que resultem em alguma informação quantitativa de interesse ou que sejam básicas para discriminação entre classes de objetos. No caso das imagens mamográficas, a forma e a textura são características poderosas que auxiliam na diferenciação entre nódulos benignos ou malignos e até mesmo no caso de tecidos saudáveis.

O último estágio envolve o reconhecimento. Reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo seu descritor [Gonzales e Woods, 2001]. Em nosso caso, a identificação de um nódulo, digamos benigno, requer a associação dos descritores para aquele nódulo com o rótulo benigno.

2.2 Análise de componentes principais

A análise de componentes principais (PCA, do inglês *principal components analysis*) [Hyvarinen et al., 2001] [Jain, 1989] é uma técnica estatística poderosa que pode ser utilizada para estudar correlações entre dados, ou seja, determinar as direções principais dos mesmos. Entende-se como direções principais o conjunto de vetores ortogonais sobre os quais os dados apresentam maior variância. O primeiro vetor representa a direção de máxima variância, o segundo vetor também está disposto segundo a direção de máxima variância mas sob a condição de ser ortogonal ao primeiro, e assim sucessivamente para o restante dos vetores.

Uma das principais aplicações da PCA é a redução de dimensionalidade através da eliminação das variáveis originais de menor variância. Embora a variabilidade total de um sistema seja definida por n variáveis, geralmente muito desta

variabilidade pode ser explicada por um número bem menor, k, de componentes principais. Desta forma a quantidade de informação contida em k é equivalente àquela existente nas n variáveis originais.

Por isso, em muitas aplicações a PCA é utilizada como uma espécie de pré-processamento dos dados, servindo como entrada para outros modelos numéricos, tais como análise discriminante e máquinas de vetor de suporte. A vantagem, neste caso, está na redução do número de parâmetros do modelo imediatamente seguinte à PCA, melhorando o desempenho e poupando tempo de processamento.

2.2.1 Cálculo das componentes principais

Vamos considerar matrizes e vetores como letras maiúsculas e minúsculas respectivamente, ambos em negrito. Então, matematicamente, consideremos a combinação linear:

$$\mathbf{z} = \mathbf{V}^T\mathbf{x} \tag{2.1}$$

A matriz de autocovariância de $\mathbf{z}$ deve, pois, ser diagonal, de forma que:

$$\mathbf{C}_z = E\left\{\mathbf{Z}\cdot\mathbf{Z}^T\right\} = \Lambda \tag{2.2}$$

Sabe-se, porém, que:

$$\mathbf{C}_z = E\left\{\mathbf{z}\cdot\mathbf{z}^T\right\} = E\left\{\mathbf{V}^T\cdot\mathbf{z}\cdot\mathbf{z}\cdot\mathbf{V}\right\} = \mathbf{V}^T\cdot E\left\{\mathbf{x}\cdot\mathbf{x}^T\right\}\cdot\mathbf{V} = \mathbf{V}^T\cdot\mathbf{C}_x\cdot\mathbf{V} \tag{2.3}$$

das equações 2.2 e 2.3, tem-se que:

$$\mathbf{V}^T\cdot\mathbf{C}_x\cdot\mathbf{V} = \Lambda \tag{2.4}$$

Portanto, $\mathbf{V}^T$ é a matriz ortogonal com NxK elementos que diagonaliza a matriz $\mathbf{C}_x$. Como resultado clássico da álgebra, $\mathbf{V}^T$ é a matriz cujas linhas são os autovetores da matriz de $\mathbf{C}_x$, correspondentes aos autovalores em ordem crescente de variância. Λ é uma matriz diagonal NxN cujos elementos são os autovalores de $\mathbf{C}_x$, ou correspondentemente, as variâncias de $\mathbf{z}$, em ordem decrescente de energia.

2.3 Análise de componentes independentes

A análise de componentes independentes (ICA, do inglês *independent component analysis*) é um método visto como uma extensão da análise de componentes principais (PCA), já que para obter independência, garantimos a descorrelação, que é um resultado da PCA. A ICA foi desenvolvida no contexto de separação cega de fontes (BSS, *blind source separation*), em que o problema é definido na estimação da saída de uma fonte conhecida, quando esta fonte recebe vários sinais misturados e desconhecidos. ICA tem sido aplicada em diversas áreas, como por exemplo: áudio, radar, instrumentação médica, comunicação móvel, engenharia biomédica e outras.

A BSS representa um grande problema na engenharia, pois a técnica mais utilizada anteriormente era a PCA, que utiliza apenas estatísticas de segunda ordem, o suficiente apenas para descorrelacionar um conjunto de dados, mas não necessário para independência, que requer estatística de alta ordem. Por esta razão a ICA é vista como um método mais "robusto"que PCA, pois se PCA consegue descorrelacionar as fontes não observáveis, ICA consegue deixá-las mutualmente estatisticamente independentes.

2.3.1 Definições

Considere que sejam observadas n misturas lineares $x_1, ..., x_n$, modeladas como combinação linear de n funções de base

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n, \forall i = 1, ..., n \tag{2.5}$$

e que cada mistura x_i, assim como cada componente independente $s_1, ..., s_n$ seja uma variável aleatória e a_{ij} os coeficientes (pesos) da mistura linear.

Assume-se que tanto as variáveis da mistura quanto aquelas das componentes independentes têm média zero. Por conveniência, será usada a notação vetorial em vez de somas, como aquelas vistas na equação 2.5, utilizando letras minúsculas e maiúsculas, para representar, respectivamente, vetores e matrizes. Dessa maneira, podemos reescrever a Equação anterior da seguinte forma:

$$\mathbf{x} = \mathbf{As} \tag{2.6}$$

O objetivo da técnica é recuperar as fontes $\mathbf{s}$, através de $\mathbf{x}$, sem nenhuma informação sobre as propriedades de $\mathbf{A}$.

O modelo estatístico definido na Equação 2.6 é chamado de modelo de análise de componentes independentes. Esse modelo descreve os dados observados pelo processo de mistura das componentes independentes s_j, que não podem ser observadas diretamente. é preciso estimar tanto $\mathbf{s}$ quanto a matriz de mistura $\mathbf{A}$, que também é desconhecida, pois tudo o que se observa é o vetor $\mathbf{x}$.

O problema do modelo de dados de ICA é estimar a matriz $\mathbf{A}$ usando apenas a informação contida na matriz $\mathbf{x}$. Para tanto, é preciso fazer suposições tão gerais quanto possível [Hyvarinen et al., 2001]. Portanto, supõe-se que:

- As componentes $s_1, ..., s_n$ são estatisticamente independentes;
- As componentes têm distribuições não-gaussianas;

2.3.2 Definição de independência

Sejam y_1 e y_2 duas variáveis aleatória, elas serão ditas independentes se a ocorrência ou não ocorrência de y_1 não influenciar na ocorrência ou não ocorrência de y_2, e vice-versa. Matematicamente, independência estatística é definida em termos da densidade de probabilidade. As variáveis y_1 e y_2 são ditas independentes se e somente se

$$p\left(y_1, y_2\right) = p_1\left(y_1\right) p_2\left(y_2\right). \tag{2.7}$$

Em palavras, a densidade conjunta $p_{y_1,y_2}(y_1, y_2)$ de y_1 e y_2 deve ser fatorada nos produtos das densidades marginais $p_{y_1}(y_1)$ e $p_{y_2}(y_2)$. Se duas variáveis são independentes, também são descorrelacionadas, mas o contrário não é verdadeiro. Duas variáveis aleatórias serão descorrelacionadas se a covariância c_{y_1,y_2} é zero:

$$c_{y_1,y_2} = E\left\{\left(y_1 - m_{y_1}\right)\left(y_2 - m_{y_2}\right)\right\} = 0 \tag{2.8}$$

ou equivalentemente,

$$r_{y_1,y_2} = E\left\{y_1, y_2\right\} = E\left\{y_1\right\} E\left\{y_2\right\}. \tag{2.9}$$

2.3.3 Técnicas de estimação das componentes

A não-gaussianidade é um elemento chave para a estimação do modelo de ICA, pois a matriz $\mathbf{A}$ não é identificável quando mais de uma das componentes independentes têm distribuição gaussiana. Consideremos que $\mathbf{x}$ é distribuído de acordo com o modelo de ICA na Equação 2.6, e que todas as componentes independentes têm distribuições iguais. Para estimar as componentes independentes, basta encontrar as combinações lineares de x_i, de modo que

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}. \tag{2.10}$$

Assim, podemos expressar uma combinação linear de x_i por

$$\begin{aligned} y &= \mathbf{b}^T\mathbf{x} \\ y &= \sum_i b_i x_i \\ y &= \mathbf{b}^T\mathbf{A}\mathbf{s} \end{aligned} \tag{2.11}$$

em que $\mathbf{b}$ deve ser determinado. A partir da Equação 2.11 podemos observar que y é uma combinação linear de s_i, com coeficientes dados por $\mathbf{q} = \mathbf{b}^T\mathbf{A}$. Logo obtemos

$$\begin{aligned} y &= \mathbf{q}^T\mathbf{s} \\ y &= \sum_i q_i s_i. \end{aligned} \tag{2.12}$$

Se $\mathbf{b}$ corresponder a uma das linhas da inversa de $\mathbf{A}$, então y será uma das componentes independentes e, nesse caso, apenas um dos elementos de $\mathbf{q}$ será igual a 1, enquanto todos os outros serão iguais a zero. Não é possível determinar $\mathbf{b}$ exatamente, mas podemos estimar seu valor com boa aproximação.

Uma forma de determinar $\mathbf{b}$ é variar os coeficientes em $\mathbf{q}$ e então verificar como a distribuição de $y = \mathbf{q}^T\mathbf{s}$ muda. Já que, conforme o Teorema do Limite Central [Papoulis, 2002], a soma de duas variáveis aleatórias independentes é mais gaussiana que as variáveis originais, $\mathbf{y} = \mathbf{q}^T\mathbf{s}$ normalmente é mais gaussiana que qualquer uma das s_i e menos gaussiana quando se iguala a uma das s_i. Nesse caso, apenas um dos elementos q_i de $\mathbf{q}$ é diferente de zero [Hyvarinen et al., 2001].

Como, na prática, os valores de $\mathbf{q}$ são desconhecidos e sabemos, através das Equações 2.11 e 2.12, que

$$\mathbf{b}^T\mathbf{x} = \mathbf{q}^T\mathbf{s} \tag{2.13}$$

podemos variar $\mathbf{b}$ e observar a distribuição de $\mathbf{b}^T\mathbf{x}$. Portanto, podemos tomar, como b, um vetor que maximiza a não-gaussianidade de $\mathbf{b}^T\mathbf{x}$, sendo que esse vetor necessariamente corresponde a $\mathbf{q} = \mathbf{A}^T\mathbf{s}$, vetor esse que possui apenas uma de suas componentes diferente de zero. Isso significa que y na Equação 2.11 é igual a uma das componentes independentes. Logo, a maximização da não gaussianidade de $\mathbf{b}^T\mathbf{x}$ permite encontrar uma das componentes.

2.3.4 Negentropia como medida de não-gaussianidade

Uma medida importante de não-gaussianidade é a negentropia. A definição de entropia [Hyvarinen et al., 2001] [Papoulis, 2002] pode ser generalizada para vetores de variáveis aleatórias contínuas, vindo a ser chamada entropia diferencial. A entropia de uma variável aleatória está relacionada à quantidade de informação que essa variável contém. A entropia será maior quanto mais imprevisível for a variável. Tomando uma variável aleatória $\mathbf{y}$ cuja função densidade de probabilidade é $f(\mathbf{y})$, temos a entropia diferencial dada por

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) dy. \tag{2.14}$$

Como um dos resultados fundamentais da Teoria da Informação, sabe-se que uma variável gaussiana tem a maior entropia entre todas as variáveis aleatórias de igual variância [Hyvarinen et al., 2001] [Papoulis, 2002]. Isso quer dizer que uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Essa medida é chamada negentropia, definida por

$$J(\mathbf{y}) = H\left(\mathbf{y}_{gauss}\right) - H(\mathbf{y}) \tag{2.15}$$

em que $\mathbf{y}_{gauss}$ é uma variável aleatória de mesma matriz de covariância que $\mathbf{y}$. A negentropia é sempre não-negativa, tem valor igual a zero se e somente se $\mathbf{y}$ tem distribuição gaussiana e é invariante para transformações lineares inversíveis.

Em contraste às suas qualidades como medida de não-gaussianidade, a negentropia é de difícil estimação. Por isso, é necessária a utilização de aproximações usando, por exemplo, momentos de alta ordem, como, por exemplo

$$J(\mathbf{y}) \approx \frac{1}{12} E\left\{y^3\right\}^2 + \frac{1}{48} kurt\,(y)^2 \tag{2.16}$$

onde $kurt(y)$, a kurtose de y, definida como o momento de quarta ordem da variável aleatória y, expresso por

$$kurt\,(y) = E\left\{y^4\right\} - 3\left(E\left\{y^2\right\}\right)^2. \tag{2.17}$$

A kurtose é zero para variáveis gausssianas e maior que zero para a maioria das variáveis aleatórias não-gaussianas.

2.4 Análise discriminante linear

Discriminante Linear, como o nome sugere, busca por uma combinação linear de variáveis de entrada que podem proporcionar uma separação adequada para as classes dadas. Ao invés de olhar para uma determinada forma de distribuição paramétrica, LDA utiliza uma aproximação empírica para definir um plano de decisão linear no espaço dos atributos, i.e. modelos de superfícies. A função discriminante usada pelo LDA é construída como uma combinação linear das variáveis que tentam de alguma forma maximizar as diferenças entre as classes [Lachenbruch, 1975].

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta^T x \tag{2.18}$$

Então o problema é reduzido a encontrar um bom vetor β. Existem várias variações dessa idéia, uma das mais bem sucedidas é a Regra do Discriminante Linear de Fisher. A Regra de Fisher é considerada uma classificação "sensível"no sentido de que é intuitivamente atraente. Ela faz uso do fato de que as distribuições com uma maior variância entre as classes do que dentro de cada classe é mais fácil de separar. Por isso, ela procura por uma função linear no espaço de atributos que maximiza a razão entre-grupos da soma dos quadrados (B) e o intra-grupos da soma dos quadrados (W). Isso pode ser conseguido através da maximização de

$$\frac{\beta B \beta}{\beta W \beta} \tag{2.19}$$

e o vetor que maximiza esta razão é o autovetor correspondente ao maior autovalor de $W^{-1}B$, isto é, a função discriminante linear y é equivalente a primeira variável canônica. Daí a regra do discriminante pode ser escrita como

$$x \in i \text{ se } \left|\beta^T x - \beta^T u_i\right| < \left|\beta^T x - \beta^T u_j\right|, \text{para todo } j \neq i \qquad (2.20)$$

onde $W = \sum n_i S_i$ e $B = \sum n_i \left(x_i - x\right)\left(x_i - x\right)'$ e n_i é o tamanho da amostra da classe i, S_i é a matriz de covariância da classe i, x_i é valor médio da amostra da classe i e x é a média populacional.

2.5 Máquina de vetor de suporte

A Máquina de Vetor de Suporte (SVM, do inglês, Support Vector Machine) introduzida por V. Vapnik [Vapnik, 1998] em 1995 é um método para estimar uma função para classificar os dados em duas classes [Burges, 1998]. A idéia básica do SVM está em construir um hiperplano como uma superfície de decisão de tal forma que a margem de separação entre exemplos positivo e negativos seja máxima. O termo SVM vem do fato de que os pontos do conjunto de treinamento que são mais próximos da superfície de decisão são chamados de vetores de suporte. O SVM realiza isso pelo princípio da minimização estrutural de risco que se baseia no fato de que a taxa de erro de uma máquina de aprendizagem no conjunto de teste é limitada pelo somatório taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis (V-C).

O processo começa com um conjunto de treino de pontos $x_i \in \Re^n$, $i = 1, 2, \cdots, n$ onde cada ponto x_i pertence a uma das duas classes identificadas pela rótulo $y_i \in \{-1, 1\}$. A meta da margem máxima de classificação é separar as duas classes por um hiperplano tal que a distância para os vetores de suporte é maximizada. A construção pode ser pensada da seguinte forma: cada ponto x no espaço de entrada é mapeado em um ponto $z = \Phi(x)$, de um espaço dimensional maior, chamado espaço de característica, onde os dados são separados linearmente por um hiperplano. A natureza dos dados determina como o método se desenvolve. Existem dados que são separáveis linearmente, separáveis não-lineamente e impossíveis de separar. A propriedade chave desta construção é que pode-se escrever nossa função de decisão usando uma função núcleo (*kernel*)

$K(x, y)$, que é dada pela função $\Phi(x)$ que mapeia o espaço de entrada sobre espaço de características. Essa superfície de decisão tem a equação:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b \tag{2.21}$$

onde $K(x, x_i) = \Phi(x).\Phi(x_i)$, e os coeficientes α_i e b são as soluções de um problema de otimização. Especificamente

$$\begin{gathered} \min_{w,b,\xi} \ \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{l} \xi_i \\ \text{para} \ \ y_i \left[w^T \cdot \phi\left(x_i\right) + b \right] \geq 1 - \xi_i \\ \xi_i \geq 0. \end{gathered} \tag{2.22}$$

onde $C > 0$ é um parâmetro a ser escolhido pelo usuário, que corresponde ao controle do compromisso entre a complexidade da máquina e o número de pontos não separáveis, os ξ_i são as variáveis soltas que penalizam os erros de treinamento e o b é o *bias*. A Equação 2.22 pode ser resolvida ao encontrar os multiplicadores α_i ótimos que satisfaçam a Equação 2.23

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \ \ 0 < \alpha_i < C \tag{2.23}$$

A classificação de um novo ponto de dados x é realizada pela cálculo do sinal do lado direito da Equação 2.21. Uma importante família de funções de kernel é a função de base radial (RBF, do inglês *Radial Basis Function*), mais comumente utilizado para problemas de reconhecimento de padrões, e também é utilizada neste trabalho. A RBF é definida por:

$$K(x, y) = e^{-\gamma \|x-y\|^2} \tag{2.24}$$

onde $\gamma > 0$ é um parâmetro que também é definido pelo usuário.

Capítulo 3

Materiais e métodos

3.1 Metodologia

A metodologia proposta neste trabalho tem a intenção de realizar dois tipos de classificação dos tecidos mamário em mamografias digitais:

- classificação em nódulos ou não-nódulos;
- classificação em benignos ou malignos.

O método é baseado em quatro passos: aquisição da imagem, redução de dimensionalidade, extração de características e classificação. Todos estes passos são encontrados na Figura 3.1 e serão descritos neste capítulo.

3.1.1 Aquisição de imagens

A aquisição da imagem é o passo onde é realizada a obtenção das mamografias e a seleção manual das regiões correspondentes ao tecido com e sem nódulo. A Figura 3.2 mostra um exemplo de seleção manual da ROI. Em seguida, para que todas as regiões de interesse tenham o mesmo tamanho, elas foram redimensionadas para 32x32 pixels. Após isto, uma equalização do histograma é realizada em cada uma das ROIs para enfatizar características não visíveis anteriormente, conforme podemos verificar na figura 3.3.

Para o desenvolvimento e avaliação da metodologia proposta, nós usamos duas bases de dados de mamografias disponíveis publicamente: a *Digital*

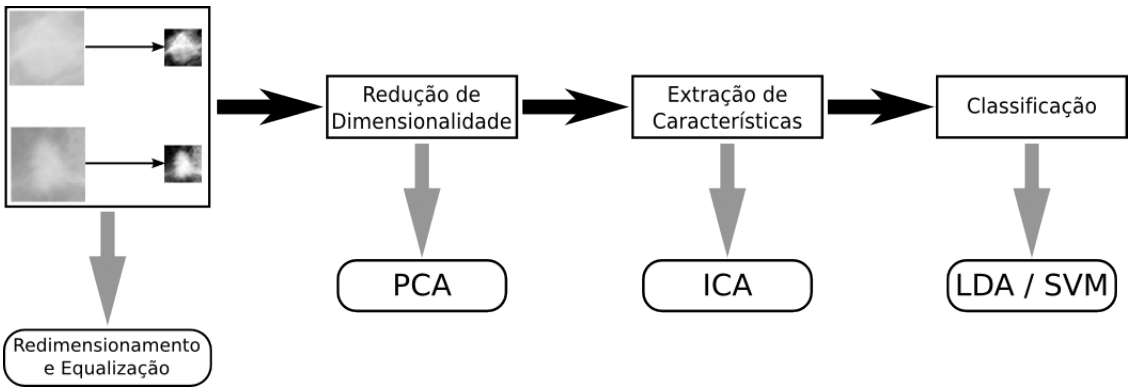

Figura 3.1: Metodologia proposta em quatro passos: aquisição de imagens, redução de dimensionalidade usando PCA, extração de características por ICA e classificação através de LDA e SVM.

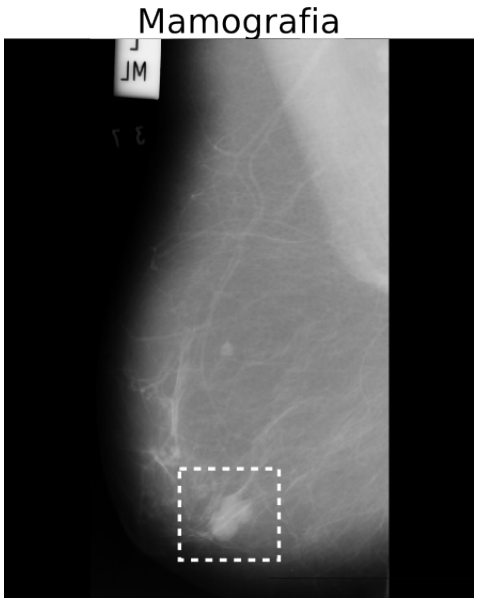

Figura 3.2: Exemplo de seleção da região de interesse. Fonte: Banco de Dados MIAS [Suckling et al., 1994]. Identificação mdb0005.pgm.

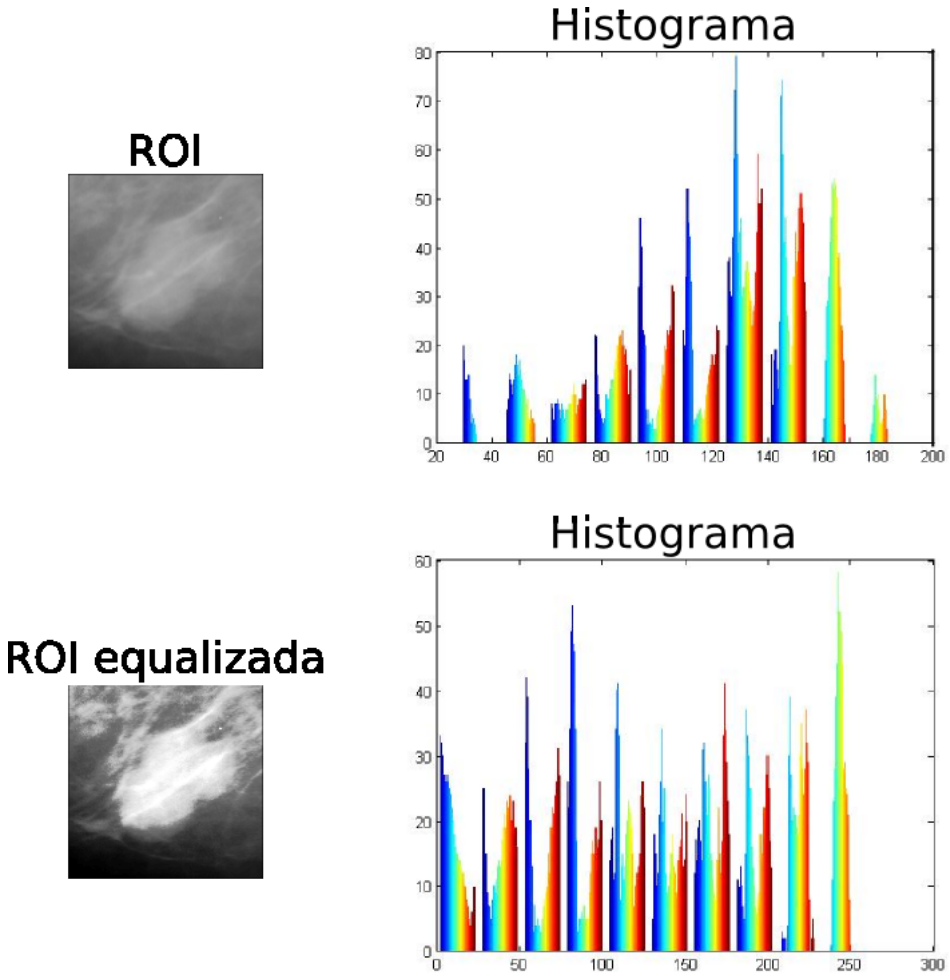

Figura 3.3: Região de interesse da Figura 3.2 antes e depois da equalização do histograma e seus respectivos histogramas.

Database for Screening Mammography (DDSM) [Heath et al., 1998] e mini-MIAS (*Mammographic Image Analysis Society*) *database* [Suckling et al., 1994].

A base de dados DDSM contêm 2620 casos com 2 tipos de incidências padrão (médio-lateral obliquo e crânio-caudal) de ambas as mamas, adquiridas do Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital e Washington University na St. Louis School of Medicine. Os dados compreendem estudos de pacientes de diferentes etinias. A DDSM contém descrições das lesões mamográficas em termos de imagens mamográficas da American College of Radiology chamada de Breast Imaging Reporting and Data System (BI-RADS) [Heath et al., 1998]. Mamogramas da base de dados DDSM foram digitalizadas por diferentes scanners dependendo da fonte dos dados de cada instituição e tem resolução entre 42 e 50 microns. Dos 2620 casos, foram selecionados 3600 regiões de interesse, onde 900 são benignas, 900 são malignas e 1800 são normais.

A base de dados mini-MIAS foi fornecida pela *Mammografic Institute Analysis Society* (MIAS) [Suckling et al., 1994]. As mamografias têm um tamanho de 1024x1024 pixels e resolução de 200 micron. Esta base de dados é composta por 332 mamogramas, tanto de mamas do lado direito quanto do lado esquerdo, de 161 pacientes, onde 53 são diagnosticados como malignos, 69 benignos e 206 normais. As anormalidades são classificadas pelo tipo de anormalidade encontrada (calcificação, massas circunscritas, distorções assimétricas da arquitetura e outros). Nós selecionamos desta base de dados 200 regiões de interesse, onde 50 são benignas, 50 malignas e 100 normais.

Todos os casos de regiões sem massa, de ambas as base de dados, foram retiradas de casos que não têm região de massa. Isto é, não foram retiradas regiões normais de casos que continham qualquer anormalidade.

3.1.2 Redução de dimensionalidade

O segundo passo é utilizar a análise de componentes principais para realizar a redução de dimensionalidade. Conseguimos isso ao usarmos apenas as k componentes principais de maior variância, reduzindo desta forma informações redundantes nas imagens. A Figura 3.4 ilustra as 50 componentes principais de maior variância extraídas das ROIs de mamografias que fazem parte do conjunto

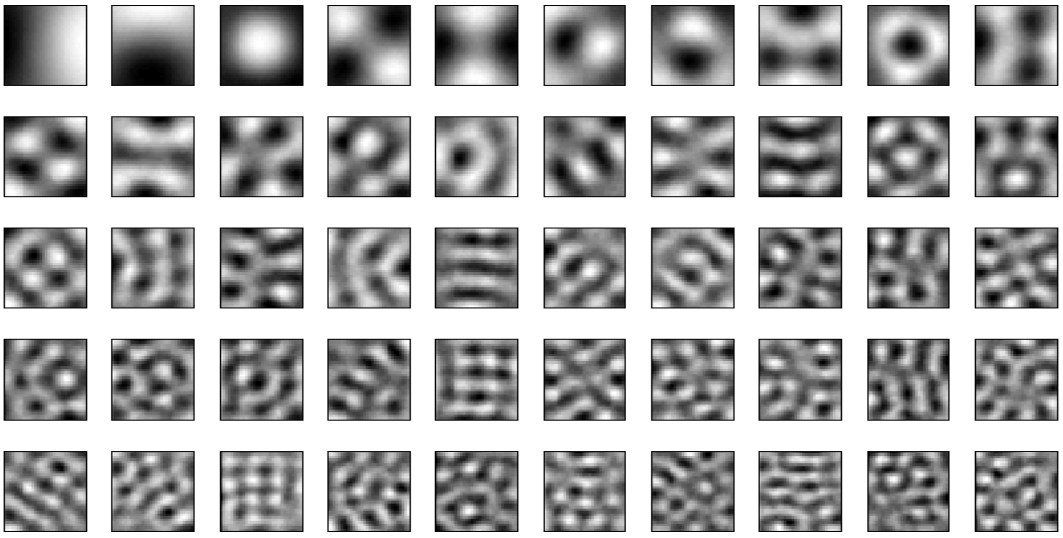

Figura 3.4: As 50 primeiras componentes principais de ROIs de mamografias, ordenadas por variância.

de treinamento da base de dados DDSM.

Em seguida, projetamos cada mamografia do treinamento no espaço das k componentes selecionadas para obter uma nova representação destas imagens. Esta projeção é dada pela Equação 3.1.

$$\mathbf{z} = \mathbf{V}^T\mathbf{x} \tag{3.1}$$

onde V^T é a matriz ortogonal com NxK elementos que diagonaliza a matriz C_x. Cada linha de V^T é uma componente principal e está ordenada por variância, a primeira componente é a de maior variância e a k-ézima componente é a que tem a menor variância neste conjunto. Desta forma, podemos obter um novo conjunto de treinamento, que terá uma dimensão reduzida em relação ao conjunto de treinamento original. Enfatizamos ainda que o conjunto de teste não foi utilizado para encontrar as componentes principais. O grupo de treinamento será utilizado no passo seguinte, a extração de característica.

3.1.3 Extração de características

Considerando que uma imagem é formada pela combinação linear de n imagens base, tal qual a Figura 3.5. O terceiro passo é obter um conjunto de imagens

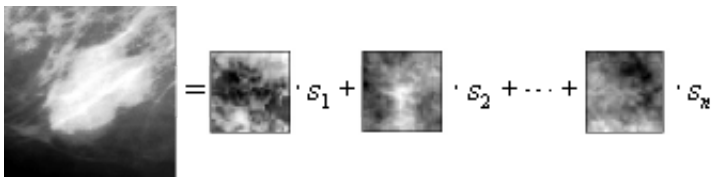

Figura 3.5: Região de Interesse como uma combinacao linear de suas características. Semelhante a Equação 2.5

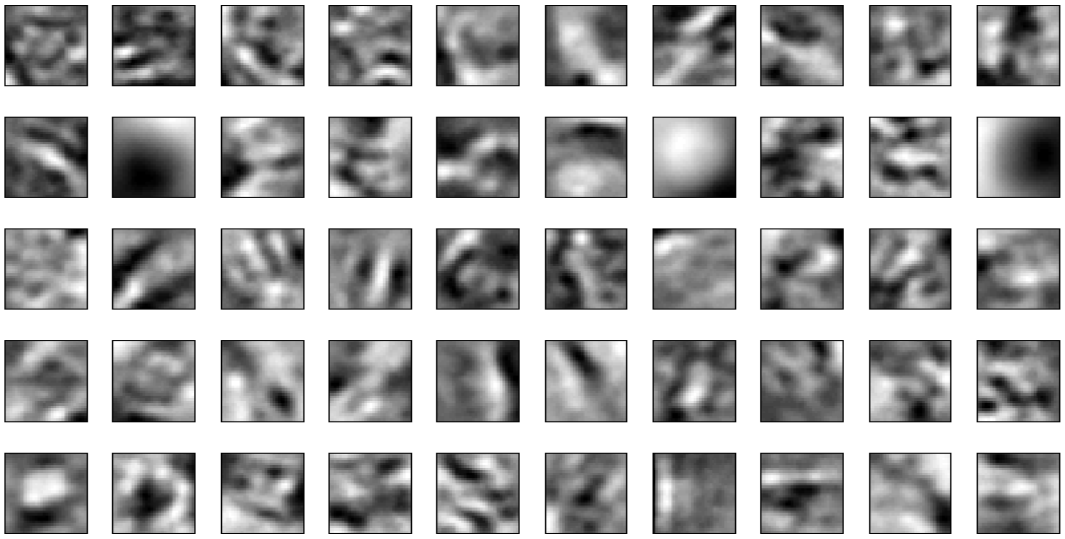

Figura 3.6: As 50 imagens base de ROIs de mamografias, após o pré-processamento de PCA para redução de dimensionalidade.

bases, utilizando análise de componentes independentes (ICA), para a extração de características.

A Figura 3.6 ilustra 50 imagens base de mamografias de z, que são as ROIs de treinamento da base de dados DDSM projetadas no espaço das componentes principais.

Lembrando novamente que o conjunto de teste não fez parte da estimação das imagens base, apenas o conjunto de treinamento com a dimensionalidade reduzida. Em seguida projetamos cada mamografia no espaço das imagens bases para obter uma nova representação destas imagens. Esta projeção é dada de forma análoga à projeção feita pelo PCA, conforme podemos ver na Equação 3.2.

$$\hat{\boldsymbol{x}} = \mathbf{A}^{\boldsymbol{T}}\mathbf{z} \tag{3.2}$$

3.1.4 Classificação

No último passo, usamos dois classificadores: análise discriminante linear (LDA) e máquinas de vetores de suporte (SVM) para classificar os tecidos em nódulos ou não-nódulos e em benignos ou malignos. Ambos os classificadores foram descritos no capítulo anterior.

Para fins de comparação, os dois classificadores receberam o mesmo conjunto de treinamento e teste, escolhidos entre as ROIs de forma aleatoriamente. No capítulo seguinte observaremos os resultados obtidos pelos dois classificadores para cada um dos casos.

3.2 Validação do método de classificação

Para avaliar o classificador em relação à sua capacidade de diferenciação, analisamos a sua sensibilidade, especificidade e acurácia. Para entendermos melhor o que cada um destes termos significa, vamos definir as variáveis que serão utilizadas nas suas definições.

- Verdadeiro Positivo (VP) - Diagnóstico do nódulo classificado corretamente como um nódulo;
- Falso Positivo (FP) - Diagnóstico do não-nódulo classificado erroneamente como um nódulo;
- Verdadeiro Negativo (VN) - Diagnóstico de não-nódulo classificado corretamente como um não-nódulo;
- Falso Positivo (FN) - Diagnóstico do nódulo classificado erroneamente como um não-nódulo.

A sensibilidade indica quão bom é o teste para identificar a patologia e é definida por:

$$sensibilidade = \frac{VP}{(VP + FN)} \tag{3.3}$$

A especificidade indica quão bom é o teste para identificar pacientes sem patologias e é definida por:

$$especificidade = \frac{VN}{(VN + FP)} \tag{3.4}$$

A acurácia é a taxa de sucesso ou acerto do teste e é dada por:

$$acuracia = \frac{(VP + TN)}{(VP + TN + FP + FN)} \tag{3.5}$$

Os cálculos para encontrar sensibilidade, especificidade e acurácia na classificação entre benignos e malignos é dada de forma análoga.

Capítulo 4

Resultados

Aqui são descritos os resultados obtidos usando o método proposto no capítulo anterior.

4.1 Base de dados das mamografias

Dos 2620 casos da base de dados DDSM, foram selecionadas 3600 regiões de interesse, onde 900 são benignos, 900 são malignas e 1800 são normais e da base de dados mini-MIAS, foram selecionadas 200 regiões de interesse, onde 50 são benignas, 50 malignas e 100 normais. Para ambas bases de dados, usamos metade das amostras para treinamento e a outra metade para testes.

Para que todas as imagens pudessem ter o mesmo tamanho, realizamos um redimensionamento nas mamografias para que elas tivessem 32x32 pixels. Em seguida transformamos esta imagem em um vetor linha de 1x1024 pixels. A figura 4.1 ilustra como este procedimento é realizado.

4.2 Aplicação da PCA / ICA

Para realizar a redução de dimensionalidade das imagens, realizamos um passo de pré-processamento explicado anteriormente. As componentes principais foram selecionadas pelos autovetores da matriz de covariância do conjunto de treinamento. Aplicando o algoritmo do FastICA [Hyvarinen et al., 2001] na matriz de treinamento, nós obtemos as imagens bases A, onde contêm as

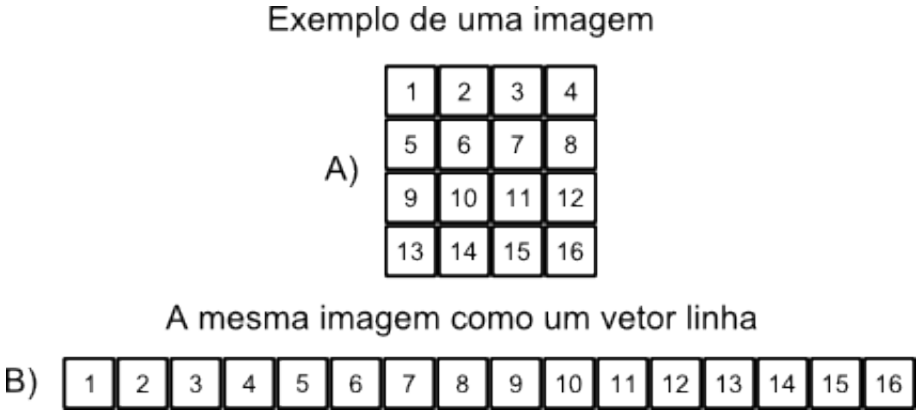

Figura 4.1: Exemplo de uma imagem transformando-se em um vetor linha. Em A) temos um exemplo de imagem onde cada quadrado representa um pixel. Em B) verificamos esta mesma imagem como um vetor linha. A numeração dentro do pixel nos ajuda a entender a posição de cada pixel após a transformação.

características das amostras. A ICA foi executada sucessivamente nos 5, 10, 20, 30, 40 e 50 autovetores mais significantes.

Em seguida projetamos todas as ROIs no espaço das imagens bases, conforme explicado na metodologia. As imagens do conjunto de treino projetadas neste espaço serão os parâmetros de entrada para o treinamento do classificador e as imagens do conjunto de teste, que não fizeram parte do processo de PCA nem do processo de ICA, projetadas no espaço das imagens bases são os parâmetros de testes.

4.3 Aplicação da LDA

A classificação de nódulo e não-nódulo e benigno ou maligno foi dada através da ferramenta estatística multivariada chamada de análise discriminante, que visa gerar critérios para separar observações em grupos através de funções das variáveis associadas a estas observações [Huberty, 1994]. Em nosso caso, cado observação é uma ROI.

A Tabela 4.1 mostra os resultados obtidos na classificação entre nódulo e não-nódulo. Este experimento obteve uma acurácia de até 90,11% para amostras da base de dados DDSM com 30 componentes e para a base de dados mini-MIAS nos

Base de Dados	Componentes	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
DDSM	5	74	87,77	80,88
	10	74	87,77	80,88
	20	78	88,22	83,11
	30	**82,22**	**98,00**	**90,11**
	40	82,22	95,77	88,99
	50	80,66	98,00	89,33
MIAS	5	76,00	80,00	78,00
	10	76,00	82,00	79,00
	20	76,00	82,00	79,00
	30	**80,00**	**90,00**	**85,00**
	40	78,00	88,00	83,00
	50	78,00	88,00	83,00

Tabela 4.1: Resultados da execução do ICA com LDA para classificação de ROIs em nódulo ou não-nódulo. Em negrito podemos ver o melhor resultado deste tipo de classificação para cada uma das bases de dados.

obtemos até 85% de acurácia com a mesma quantidade de componentes.

A Tabela 4.2 apresenta os resultados obtidos na classificação entre benignos e malignos. Este experimento obteve, no melhor caso, acurácia de até 95,38% para amostra DDSM com 50 componentes. Para amostras da base de dados mini-MIAS obtivemos 92% de acurácia com as mesmas 50 componentes.

4.4 Aplicação da SVM

Uma biblioteca para máquinas de vetores de suporte, chamada LIBSVM [Chang e Lin, 2003], foi usada para o treinamento e teste. Utilizamos como núcleo, a função de base radial (RBF).

Na Tabela 4.3 mostramos os resultados obtidos para a classificação entre nódulo ou não-nódulo. Neste experimento obtemos até 99,55% de acurácia para amostras DDSM e 98% de acurácia nas amostras da base de dados mini-MIAS. Ambos resultados usando 5 componentes.

A Tabela 4.4 apresenta os resultados obtidos na classificação entre benigno

Base de Dados	Componentes	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
DDSM	5	82,88	84,88	83,88
	10	94,88	92,00	93,44
	20	92,88	93,11	92,99
	30	94,22	94,44	94,33
	40	94,00	94,44	94,22
	50	**94,44**	**96,33**	**95,38**
MIAS	5	86,00	82,00	84,00
	10	86,00	92,00	89,00
	20	88,00	90,00	89,00
	30	90,00	90,00	90,00
	40	92,00	90,00	91,00
	50	**92,00**	**92,00**	**92,00**

Tabela 4.2: Resultados da execução do ICA com LDA para classificação de nódulos em benignos ou malignos. Em negrito podemos ver o melhor resultado deste tipo de classificação para cada uma das bases de dados.

Base de Dados	Componentes	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
DDSM	5	90,00	90,00	90,00
	10	92,88	94,00	93,44
	20	97,77	98,44	98,10
	30	98,66	98,88	98,77
	40	99,33	99,55	99,44
	50	**99,55**	**99,55**	**99,55**
MIAS	5	92,00	86,00	89,00
	10	92,00	88,00	90,00
	20	92,00	90,00	91,00
	30	94,00	92,00	93,00
	40	94,00	94,00	94,00
	50	**96,00**	**100,00**	**98,00**

Tabela 4.3: Resultados da execução do ICA com SVM para classificação de ROIs em nódulo ou não-nódulo. Em negrito podemos ver o melhor resultado deste tipo de classificação para cada uma das bases de dados.

Base de Dados	Componentes	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
DDSM	5	90,00	90,22	90,11
	10	94,00	94,44	94,22
	20	97,55	97,77	97,66
	30	98,44	98,00	98,22
	40	98,66	98,88	98,77
	50	**99,33**	**99,77**	**99,55**
MIAS	5	96,00	90,00	93,00
	10	98,00	90,00	94,00
	20	100,00	90,00	95,00
	30	98,00	96,00	97,00
	40	**100,00**	**100,00**	**100,00**
	50	100,00	98,00	99,00

Tabela 4.4: Resultados da execução do ICA com SVM para classificação de nódulos em benignos ou malignos. Em negrito podemos ver o melhor resultado deste tipo de classificação para cada uma das bases de dados.

e maligno. Este experimento resultou, no melhor caso, acurácia de 99,5% para amostras DDSM com 50 componentes. Para amostras da mini-MIAS nós obtivemos até 100% de acurácia com 40 componentes.

Capítulo 5

Discussão

Observando os resultados, percebemos que a SVM obteve melhores resultados porque a separação não-linear dos dados se adapta melhor ao nosso caso do que a separação baseada na técnica da maximização da variância entre os dados utilizada pela LDA. Nas figuras 5.1 e 5.2 observamos um exemplo de classificação utilizando a SVM e a LDA respectivamente.

Apesar da separação da densidade de probabilidade das classes, realizada pela LDA, estarem bem definidas, ainda existe uma pequena intersecção entre as classes, fazendo com que a SVM obtenha melhores resultados, pois procura o melhor hiperplano no espaço de características que divide estas duas classes através da minimização de risco.

Um outro fator interessante nestes resultados é que nem sempre a melhor acurácia é a de maior componente. Suspeitamos que isto acontece quando utilizamos muita informação para classificar, gerando redundâncias e confundindo o classificador, consequentemente, diminuindo a taxa de acerto. Acreditamos que o número ideal de componentes esteja entre 30 e 50, pois testes realizados com mais componentes não obteve resultados melhores, apesar de continuar obtendo resultados próximas da média.

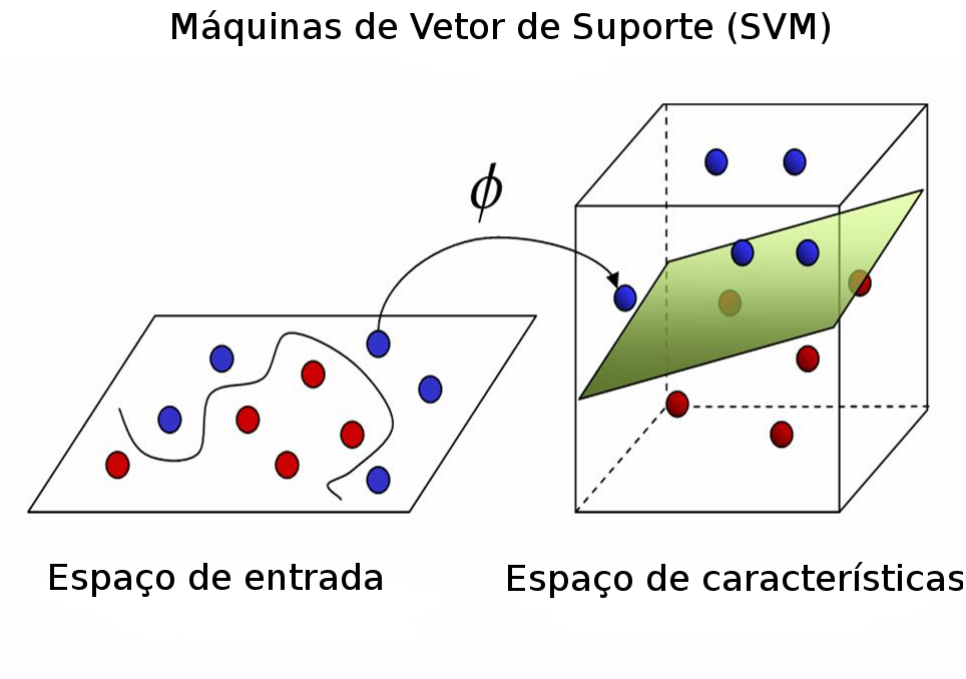

Figura 5.1: Exemplo de comportamento dos dados utilizando a SVM.

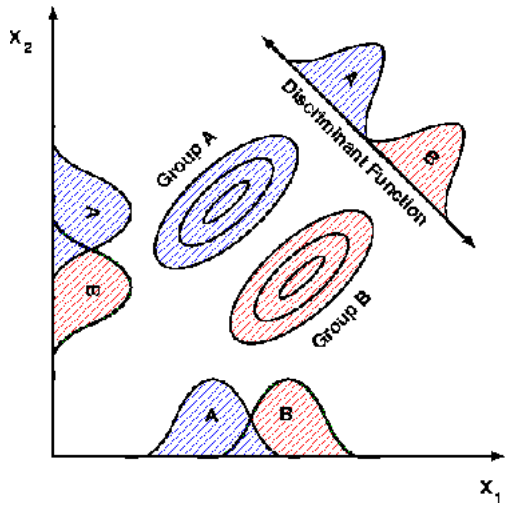

Figura 5.2: Exemplo de comportamento dos dados utilizando a LDA.

Capítulo 6

Conclusão

Neste trabalho, nós apresentamos um sistema de diagnóstico auxiliado por computador no problema de reconhecimento de câncer de mama em imagens de mamografias digitais. Classificamos pequenas regiões de interesse em nódulos ou não-nódulos e em benignos ou malignos.

Como resultados, obtemos uma taxa de acerto de até 90,11% e 85% para classificação em nódulo ou não-nódulo usando LDA nas bases de dados DDSM e mini-MIAS respectivamente. Já o SVM obteve 99,55% e 98%. No caso da classificação dos nódulos em benignos ou malignos obtemos uma acurácia de 95,38% e 92% usando LDA nas bases de dados DDSM e mini-MIAS respectivamente. Já o SVM alcançou taxas de até 99,55% e 100,00%. Baseado nestes resultados, nós concluímos que o LDA e o SVM com as bases de ICA tem um alto desempenho de predição e isto fornece um suporte significante para uma investigação clínica mais detalhada.

Em trabalhos futuros pretendemos:

- Testar a eficácia do método com uma base de dados regional, ainda em fase de construção.
- Elaborar um programa de baixo custo computacional e simples de ser manuseado, para ser utilizado por médicos e radiologistas.

Artigos publicados pelo autor

- Daniel Duarte Costa; Lúcio Flávio Campos; Allan Kardec Barros; Aristófanes Corrêa Silva. Independent Component Analysis in Breast Tissues Mammograms Images Classification Using LDA and SVM. In: The International Special Topic Conference on Information Technology Applications in Biomedicine 2007 - ITAB 2007, 2007, Tokyo. Proceedings of the IEEE Engineering in Medicine and Biology Society, 2007.

- Cristiane C. S. da Silva; Daniel Duarte Costa; Aristófanes Corrêa Silva; Allan Kardec Barros. Diagnosis of Lung Nodule using Independent Component Analysis in Computerized Tomography Images. In: International Conference on Neural Information Processing (ICONIP), 2007, Kitakyushu. Lecture Notes in Computer Science (LNCS), 2007.

- S. Comani, D. Guilhon, P. Van Leeuwen, Daniel Duarte Costa, A.K. Barros, B. Hailer, D. Grönemeyer. Effectiveness of ICA processing for feature extraction in magnetocardiographic Signals. Biomedizinische Technik, 52: CD-ROM, 2007

Referências Bibliográficas

[ACS, 2007] American Cancer Society. *Cancer Facts & Figures 2007.* Disponível em: < *http* : *//www.cancer.org/downloads/STT/CAFF*2007*PWSecured.pdf* > acesso em: 08 de Janeiro de 2008.

[Bauer et al.,1980] BAUER, W.; IGOT, J.P.; LE, G.Y.. *Chronologic du cancer mammaire Utilisant un Modele de Croissance de Gompertz.* Ann Anat Pathol 25:39-56, 1980

[Bland e Copeland,2000] Bland, K.I.; Copeland, E.M. *A Mama: Tratamento Compreensivo das Doenças Benignas e Malignas*, Ed., Manole Ltda

[Boyd et al.,1995] Boyd, N.F.; Byng, J.W.; Jong, R.A.; Fishell, E.K.; Little, L.E.; Miller, A.B.; Lockwood, G.A.; Tritcheler, D.L.; Yaffe, M.J. *Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study.* Journal of the National Cancer Institute, v. 87, p. 670-675, 1995.

[Braz et al., 2007a] Braz JR., G., E. C. Silva, A. C. Paiva, A. C. Silva e M. Gattass. *Breast Tissues Mammograms Images Classification using Moran's Index, Geary's Coefficient and SVM.* 2007a.

[Braz et al., 2007b] Braz JR., G., E. C. Silva, A. C. Paiva e A. C. Silva. *Breast Tissues Classification Based on the Application of Geostatistical Features and Wavelet Transform.* International Special Topic Conference on Information Technology Applications in Biomedicine, ITAB 2007, 6th, p. 227-230. 2007b.

[Burges, 1998] Burges, C.J.C., it A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, 1998.

[Campanini et al, 2002] Campanini Renato, Bazzani Armando, et al. *A novel approach to mass detection in digital mammography based on Support Vector Machines (SVM).* In proceedings of the 6 International Workshop in Digital Mammography (IWDM), p. 399-401, Bremem, Germany, 2002, Springer Verlag.

[Campos et al., 2005a] Campos, L. F. A., Silva, A. C., Barros, A. K., *Diagnosis of Breast cancer in digital mammograms using independent component analysis and neural networks.* LNCS 3773/Springer-Verlag, p. 460-461, 2005.

[Campos et al., 2005b] Campos, L. F. A., Silva, A. C., Barros, A. K., *Independent component analysis and neural networks applied for classification of malign, benign and normal tissues in digital mammography.* Fifth International Workshop on Biosignal Interpretation, p. 85-88, 2005.

[Chang e Lin, 2003] Chang, C.C., Lin, C.J. *LIBSVM - a Library for Support Vector Machines*, Disponível em: < *hhtp : //www.csie.ntu.edu.tw/cjlin/libsvm/* > acesso em: 08 de Janeiro de 2008.

[Chen e Chang, 2004] Chen, Y. e C. I. Chang. *A new application of texture unit coding to mass classification for mammograms.* International Conference on Image Processing, ICIP'04. 2004 5, 3335-3338.

[Christoyianni et al., 2002] Christoyianni, I., Koutras, A., Dermatas, E., Kokkinakis, G. *Computer aided diagnosis of breast cancer in digitized mammograms.* Computerized Medical Imaging and Graphics 26, p. 309-319, 2002.

[Costa et al., 2006] *Independent Component Analysis in Breast Tissues Mammograms Images Classification Using LDA and SVM.* In: The International Special Topic Conference on Information Technology Applications in Biomedicine 2007 - ITAB 2007, 2007, Tokyo. Proceedings of the IEEE Engineering in Medicine and Biology Society, 2007.

[Feig, 2006] S.A. Feig, *4-13 Computer-aided Detection (CAD) in Mammography: Does it Help the Junior or the Senior Radiologist?*, Breast Diseases: A Year Book Quarterly Volume 16, Issue 4, January-March 2006, Pages 342-343.

[Gonzales e Woods, 2001] Gonzales, Rafael C., Woods, Richard E. *Digital Image Processing*, 2 ed., Prentice Hall, 1997.

[Heath et al., 1998] Heath, M., Bowyer, K., Kopans, D., *Current Status of The Digital Database for Screening Mammography*. Digital Mammography, Kluwer Academic Publishers, p. 457-460, 1998.

[Huberty, 1994] Huberty, C.J. it Applied Discriminant Analysis. Wiley-Interscience, 1994.

[Hyvarinen et al., 2001] Hyvarinen, A., Karhunen, J., Oja, E. *Independent Component Analysis*, Nova York, John Wiley & Sons, p. 481, 2001.

[INCA, 2008] Instituto Nacional do Câncer. *Estimativa 2008 - Incidência de Câncer no Brasil*. Disponível em: $<$ *http* : *//www.inca.gov.br/estimativa/*2008/*versaofinal.pdf* $>$ acesso em: 08 de Janeiro de 2008.

[Jain, 1989] Jain, Anil K. *Fundamentals of Digital Image Processing*, 1 ed., Prentice Hall, 1989.

[Lachenbruch, 1975] Lachenbruch, P.A., *Discriminant Analysis*. Hafner Press, New York, (1975).

[Lim e Er, 2004] Lim, W. K. e M. J. ER. Classification of Mammographic Masses using Generalized Dynamic Fuzzy Neural Networks. Medical Physics 31, 1288.

[Martins et al., 2006] Martins, L. O., A. M. dos Santos, A. C. Silva e A. C. Paiva. *Classification of Normal, Benign and Malignant Tissues using Co-Occurrence Matrix and Bayesian Neural Network in Mammographic Images*. Proceedings of the Ninth Brazilian Syposium on Neural Networks, p. 479-486.

[Masotti, 2006] *A Ranklet-Based Image Representation for Mass Classification in Digital Mammograms*. Medical Physics 33, 3951.

[Minh N. Do e Martin Vetterli, 2005] *The Contourlet Transform: An eficient directional multiresolution image representation* IEEE Trans. Image Proc., 2005.

[Moayedi et al., 2007] Moayedi, F., R. Boostani, Z. Azimifar e S. Kabeti. *A Support Vector Based Fuzzy Neural Network apporach for Mass Classification in Mammography*. Digital Signal Processing. 15th International Conference, p. 240-43, 2007.

[NCI, 2006] National Cancer Institute. *Cancer stat fact sheets: Cancer of the breast (2006)*. Disponível em: $< http : //seer.cancer.gov/statfacts/html/breast.html >$ acesso em: 08 de Janeiro de 2008.

[Oliver et al. 2006] Oliver, A., J. Marti, R. Marti, A. Bosch e J. Freixenet. *A new apporach to the classification of mammographic masses and normal breast tissue*. Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) - Volume 04, p. 707-710.

[Papoulis, 2002] Athanasios, P.; Pillai, S. Unnikrishna. *Probability, Random Variables and Stochastic Processes*, 4 ed., Nova York, McGraw-Hill, p. 852, 2002.

[Suckling et al., 1994] Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., et al, *The mammographic images analysis society digital mammogram database*. Exerpta Medical, v. 1069, p. 375-378, 1994.

[Thuler, 2003] Luiz Cláudio Thuler. *Considerações sobre a prevenção do câncer de mama feminino*. Revista Brasileira de Cancerologia, v. 49(4) p. 227-238, 2003.

[Timp et al., 2007] Timp, S., C. Varela e N. Karssemeijer. *Temporal Change Analysis for Characterization of Mass Lesion in Mammography*. Medical Imaging, IEEE Transactions on 26(7), 945-953.

[Vapnik, 1998] *Statistical Learning Theory*. Wiley New York, 1998.

[Wei et al., 1995] Wei, D., H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler e M. M. Goodsitt. *Classification of Mass and Normal Breast Tissue on Digital Mammograms: Multiresolution Texture Analysis*. Medical Physics 22, 1501. 1995.

[Zhang et al., 2002] Zhang, L., Sankar, R., Qian, W. *Advances in micro-calcification clusters detection im mammography*, Comp. in Biology & Medicine, v. 32, p. 515-528, 2002.

[Zhang et al., 2004] Zhang, P., B. Verma e K. Kumar. *A Neural Genetic algorithm for feature selection and breast abnormality classification in digital mammography*. IEEE International Joint Conference on Neural Networks, 2004. Proceedings 2007, 2303-2308.