

UNIVERSIDADE FEDERAL DO MARANHÃO  
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA  
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

**LEONARDO DE OLIVEIRA MARTINS**

**DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS ATRAVÉS DO  
ALGORITMO *GROWING NEURAL GAS* E DA FUNÇÃO *K* DE RIPLEY**

São Luís  
2007

**LEONARDO DE OLIVEIRA MARTINS**

**DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS ATRAVÉS DO  
ALGORITMO *GROWING NEURAL GAS* E DA FUNÇÃO *K* DE RIPLEY**

Dissertação apresentada ao curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica na área de Ciência da Computação.

Orientador: Prof. Dr. Aristófanés Corrêa Silva

Co-orientador: Prof. Dr. Anselmo Cardoso de Paiva

São Luís  
2007

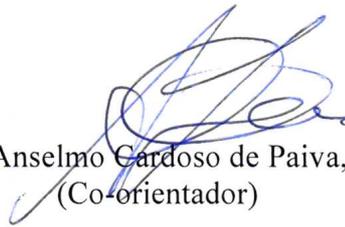
**DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS  
ATRAVÉS DO ALGORITMO GROWING NEURAL  
GAS E DA FUNÇÃO K DE RIPLEY**

**Leonardo de Oliveira Martins**

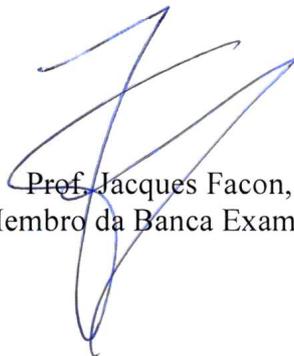
Dissertação aprovada em 07 de dezembro de 2007.



Prof. Aristófares Corrêa Silva, Dr.  
(Orientador)



Prof. Anselmo Cardoso de Paiva, Dr.  
(Co-orientador)



Prof. Jacques Facon, Dr.  
(Membro da Banca Examinadora)



Prof. Alexandre César Muniz de Oliveira, Dr.  
(Membro da Banca Examinadora)

*À minha família*

## AGRADECIMENTOS

A Deus, fonte de toda a Vida.

A meus pais, Braga e Helena, por todo o apoio e exemplo que me deram durante toda a minha vida.

A minhas irmãs, Carol e Cláudia, pelo carinho e companheirismo dedicados a mim.

Aos meus orientadores, Aristófanês Silva e Anselmo Paiva, pela paciência e dedicação imprescindíveis à realização desse trabalho.

Aos meus amigos, por participarem comigo dessa grande conquista.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro concedido ao longo do mestrado em forma de bolsa de estudo.

A todos os que direta ou indiretamente contribuíram para a realização deste trabalho.

## RESUMO

O câncer de mama apresenta-se como um grave problema de saúde pública em vários países do mundo. Sistemas de Detecção e Diagnóstico baseados em computador (CAD/CADx) vêm sendo usados com relativo sucesso no auxílio aos profissionais de saúde. O objetivo de tais sistemas não é substituir o profissional, mas unir forças com o objetivo de detectar precocemente os diferentes tipos de câncer. A principal contribuição deste trabalho é apresentar uma metodologia para detecção de massas em imagens mamográficas digitais, utilizando para tanto o algoritmo *Growing Neural Gas* para a segmentação da imagem e a função *K* de Ripley para descrever a textura dos objetos segmentados. A classificação desses objetos é feita através de uma Máquina de Vetor de Suporte (*Support Vector Machine - SVM*), a qual separa os mesmos em dois grupos: massa e não-massa. A metodologia obteve 89,30% de acerto e uma taxa de 0,93 falso-positivos por imagem.

Palavras-chave: Textura, Função *K* de Ripley, *Growing Neural Gas*, Detecção auxiliada por computador, Mamografia.

## ABSTRACT

Breast cancer is a serious public health problem in several countries of the world. Computer-Aided Detection/Diagnosis systems (CAD/CADx) have been used with relative success in aid to health care professionals. The goal of such systems is not to replace the professional, but join forces in order to early detect the different types of cancer. The main contribution of this work is to present a methodology for detecting masses in digitized mammograms using the algorithm Growing Neural Gas for the segmentation of the image and Ripley's  $K$  function to describe the texture of segmented objects. The classification of these objects is accomplished through a Support Vector Machine (SVM), which separates them into two groups: masses and non-masses. The methodology obtained 89,30% of accuracy and a rate of 0,93 false-positive per image.

Keywords: Texture, Ripley's K Function, Growing Neural Gas, Computer-aided detection, Mamography.

## ARTIGOS CIENTÍFICOS PUBLICADOS PELO AUTOR

MARTINS, Leonardo de Oliveira ; BRAZ JUNIOR, Geraldo ; SILVA, Erick Corrêa ; SILVA, Aristófanés Corrêa ; PAIVA, A. C. **Classification of Breast Tissues in Mammogram Images using Ripley's K Function and Support Vector Machine.** In: International Conference on Image Analysis and Recognition (ICIAR), 2007, Montreal. Lecture Notes in Computer Science (LNCS), 2007.

MARTINS, Leonardo de Oliveira ; SILVA, Erick Corrêa ; SILVA, Aristófanés Corrêa ; PAIVA, A. C. ; GATTASS, Marcelo . **Classification of Breast Masses in Mammogram Images using Ripley's K Function and Support Vector Machine.** In: 5<sup>th</sup> International Conference on Machine Learning and Data Mining (MLDM), 2007, Leipzig. Lecture Notes in Artificial Intelligence (LNAI), 2007.

MARTINS, Leonardo de Oliveira ; SANTOS, A. M. ; SILVA, Aristófanés Corrêa ; PAIVA, A. C. . **Classification of Normal, Benign and Malignant Tissues Using Co-occurrence Matrix and Bayesian Neural Network in Mammographic Images.** In: IX Brazilian Neural Networks Symposium, 2006, Ribeirão Preto. Anais do IX Brazilian Neural Networks Symposium, 2006.

MARTINS, Leonardo de Oliveira ; SANTOS, A. M. ; SILVA, Aristófanés Corrêa ; PAIVA, A. C. . **Classificação de tecidos normais, benignos e malignos utilizando matrizes de co-ocorrência e redes neurais Bayesianas em imagens de mamografia.** In: I Jornada de Informática do Maranhão - JIM, 2006, São Luís. Anais da I Jornada de Informática do Maranhão, 2006

## LISTA DE FIGURAS

Figura 1.1 - Esquematização do funcionamento de um software CAD típico	19
Figura 2.1 - Exemplo de mamografia	26
Figura 2.2 - Principais incidências de raios X em um exame de mamografia	28
Figura 2.3 - Diagrama de um mamógrafo digital	31
Figura 2.4 - Anormalidades encontradas em tecidos mamários	32
Figura 2.5 - Classificação das massas de acordo com o aspecto de suas bordas	33
Figura 2.6 - Classificação das massas de acordo com sua forma	33
Figura 2.7 - Passos fundamentais em processamento de imagens	37
Figura 2.8 – Representação de um histograma e seu respectivo histograma acumulado	38
Figura 2.9 – Representação de um histograma e seu respectivo histograma equalizado	39
Figura 2.10 - Ilustração do funcionamento da Função $K$ de Ripley para um dado raio $t$	44
Figura 2.11 - Ilustração do funcionamento da Função $K$ de Ripley modificada para um dado raio $t$	46
Figura 2.12 – Ilustração da medida geométrica de excentricidade	48
Figura 2.13 – Ilustração da medida geométrica de circularidade	49
Figura 2.14 – Ilustração da compacidade de dois objetos	49
Figura 2.15 – Ilustração da densidade circular de três objetos	51
Figura 2.16 – (a) Diagrama de Voronoi (linhas mais finas) e a respectiva triangulação de Delaunay (linhas mais grossas); (b) Triangulação induzida de Delaunay, com as regiões sombreadas apresentando $P(x) > 0$	56
Figura 2.17 – Evolução do algoritmo GNG	63
Figura 2.18 - Separação de duas classes através de hiperplanos	65

Figura 2.19 - Vetores de suporte (destacados por círculos)	67
Figura 3.1 – Metodologia proposta	73
Figura 3.2 – Imagem de mamografia antes da segmentação da mama (a); imagem após a segmentação (b)	75
Figura 3.3 – Imagem dividida em janelas	76
Figura 3.4 – Mamografia segmentada através do algoritmo GNG. (a) Imagem original; (b) utilizando 7 grupos; (c) utilizando 8 grupos; (d) utilizando 9 grupos	78
Figura 3.5 – Objetos considerados “massa” (a,b,c,d) e “não-massa” (e,f e g), de acordo com sua geometria	80
Figura 3.6 – Localização do centro de massa	81
Figura 3.7 – Exemplo de uma massa segmentada antes e depois do processo de equalização de histograma	82
Figura 3.8 – Resultado final do processo de detecção de massas, indicando a localização correta da massa, em vermelho	84
Figura 4.1 – Localização de uma massa. Fonte: <i>Digital Database for Screening Mammography – DDSM</i>	90
Figura 4.2 – Segmentação correta de uma massa	91
Figura 4.3 – Primeiro caso de estudo	94
Figura 4.4 – Segundo caso de estudo	96
Figura 4.5 – Terceiro caso de estudo	98

## LISTA DE TABELAS

Tabela 4.1 – Resultados para o treinamento (classificação baseada em geometria)	87
Tabela 4.2 – Resultados para os testes (classificação baseada em geometria)	87
Tabela 4.3 – Resultados para o treinamento (classificação baseada em textura)	89
Tabela 4.4 – Resultados para os testes (classificação baseada em textura)	89

## LISTA DE SIGLAS

CAD – *Computer- Aided Detection*

CADx – Computer- Aided Diagnosis

DDSM – Digital Database for Screening Mammography

GNG – Growing Neural Gas

INCA – Instituto Nacional do Câncer

MVS – Máquina de Vetor de Suporte

SBM – Sociedade Brasileira de Mastologia

## SUMÁRIO

LISTA DE FIGURAS	9
LISTA DE TABELAS	11
LISTA DE SIGLAS	12
1 INTRODUÇÃO	15
2 FUNDAMENTAÇÃO TEÓRICA	25
2.1 Mamografia	25
2.1.1 Massas e outras anormalidades	31
2.2 Processamento de imagens digitais	34
2.2.1 Passos fundamentais do processamento de imagens digitais	37
2.2.2 Equalização de histograma	37
2.2.3 Limiarização ( <i>Thresholding</i> )	40
2.3 Análise de imagens baseada em textura	41
2.3.1 Textura	41
2.3.2 Função <i>K</i> de Ripley	43
2.4 Medidas Geométricas	46
2.4.1. Excentricidade ( <i>E</i> )	47
2.4.2 Circularidade ( <i>C</i> )	48
2.4.3. Compacidade ( $C_0$ )	49
2.4.4. Desproporção Circular ( <i>D</i> )	50
2.4.5. Densidade Circular ( $D_c$ )	50
2.5 Mapas auto-organizáveis	51
2.5.1 Algoritmo <i>Growing Neural Gas</i>	53
2.6 Máquinas de Vetor de Suporte	63
2.7 Medidas de Desempenho	68
3 METODOLOGIA	70
3.1 Amostra	70
3.2 <i>Software e hardware</i> utilizados	71
3.3 Metodologia proposta	72
3.3.1 Pré-processamento	73
3.3.2 Segmentação	75
3.3.3 Redução dos candidatos a massa	79
3.3.4 Redução dos falsos positivos	80

4 RESULTADOS	85
4.1 Treinamento e teste dos classificadores MVS	85
4.1.1 Redução dos candidatos a massa	86
4.1.2 Redução dos falsos positivos	88
4.2 Resultados da detecção	89
4.3 Casos de estudo	92
4.3.1 Primeiro caso	93
4.3.2 Segundo caso	95
4.3.3 Terceiro caso	97
5 CONCLUSÃO	99
5.1 Trabalhos futuros	100
Referências	102

## 1 INTRODUÇÃO

O câncer de mama é o segundo tipo de câncer mais comum no mundo, e o mais freqüente entre as mulheres. A cada ano, no Brasil, cerca de 48.000 mulheres desenvolvem câncer de mama. No ano passado, essa doença ocasionou a morte de pelo menos 9.335 mulheres brasileiras (INCA, 2005). A estimativa para este ano é de 53 mil novos casos, incluindo de 12 a 13 mil mortes (SBM, 2007). Este tipo de câncer é o mais freqüente nas mulheres das regiões Sul (69/100.000), Centro-Oeste (38/100.000) e Nordeste (27/100.000). Na região Norte é o segundo tumor mais incidente (15/100.000) (INCA, 2005).

Em duas décadas, a taxa de mortalidade por câncer de mama no Brasil apresentou elevação de 68%, passando de 5.77 em 1979 para 9.70 mortes por 100 mil mulheres em 1998 (INCA, 2003). Dentre as explicações para esse aumento estão o envelhecimento da população, a mudança do perfil reprodutivo, a exposição a poluentes, o sedentarismo, a obesidade, dentre outros. Como mudanças no perfil reprodutivo feminino podem ser citadas a gravidez tardia, a nuliparidade, a diminuição no número de gestações e o tempo de amamentação (SILVA, 2006).

No estado do Maranhão, o câncer de mama está em segundo lugar em total de mortes por câncer, em mulheres. Entre 1979 e 1983, o câncer de mama obteve 8,32% do total de mortes por câncer. Durante o período entre 1995 e 1999, ano na qual foi realizada a última pesquisa, aumentou para 11,06% do total de mortes por câncer no estado (INCA, 2002).

Os diferentes tipos de câncer são resultados de alterações do

DNA, que levam a uma proliferação celular descontrolada. Normalmente, as células crescem e se reproduzem, dando origem a novas células necessárias para manter a saúde do organismo. Porém, por razões ainda desconhecidas, algumas vezes esse processo dá errado e as células acabam formando mais células de maneira descontrolada. As massas resultantes dessa má formação são chamados de tumores ou neoplasias, as quais são classificadas como malignas ou benignas. As neoplasias benignas têm crescimento organizado, em geral lento, e o tumor apresenta limites bem nítidos. Na neoplasia maligna, o crescimento é mais rápido, desordenado e infiltrativo. Tais massas têm capacidade de se desenvolver em outras partes do corpo, fenômeno este denominado metástase. Somente a neoplasia maligna é chamada de câncer.

Entretanto, sabe-se que as chances de cura do câncer de mama são relativamente altas se detectado nos estágios iniciais. Estima-se que a sobrevivência média geral cumulativa após cinco anos seja de 65% nos países desenvolvidos e de 56% para países em desenvolvimento. Na população mundial, esse índice é de 61% (INCA, 2005).

O desenvolvimento e a popularização de sistemas de Raio-X dedicados exclusivamente ao rastreamento do câncer de mama, nas décadas de 60 e 70, possibilitou pela primeira vez um aumento nas taxas de sobrevivência das pacientes. A mamografia, como é chamado a radiografia da mama, permite que um radiologista identifique lesões da ordem de milímetros, permitindo assim o diagnóstico precoce e possibilitando maiores chances de cura. Na Inglaterra, na década de 70, a taxa de sobrevivência 5 anos após o diagnóstico era de 50%. Hoje, graças a técnicas como a mamografia de alta resolução, essa taxa aumentou para

80% (OFFICE FOR NATIONAL STATISTICS, 2005).

Dessa maneira, a mamografia vem se mostrando a maneira mais eficaz de detectar precocemente as lesões da mama em estados iniciais de sua formação, contribuindo muitas vezes para a cura completa dessa doença. Esse exame é apontado como o principal método de detecção de neoplasias da mama em estágios iniciais, sendo capaz de detectar alterações ainda não palpáveis. Isso favorece o tratamento precoce, menos agressivo e com melhores resultados (BOYLE, 2002).

Entretanto, a mamografia não é perfeita. A sensibilidade desse exame pode variar bastante, em decorrência de fatores como qualidade do exame, experiência do radiologista, e densidade do tecido mamário. A imagem adquirida através do exame contém apenas tons de cinza. Por essa razão, em várias ocasiões não fica muito claro para o profissional visualizar a anormalidade e tampouco diagnosticá-la com exatidão. Outra dificuldade está no cansaço visual provocado pelo número elevado de exames que o radiologista geralmente examina por dia. Nesse caso, o desgaste físico pode comprometer uma série de diagnósticos.

A detecção precoce do câncer de mama obteve um avanço significativo com o advento da mamografia digital ou mamografia computadorizada. Uma das principais vantagens é que o radiologista pode armazenar a imagem em um computador e melhorá-la, alterando a orientação, brilho, e contraste, através de técnicas de processamento de imagens.

De uma maneira geral, o objetivo do uso do processamento digital de imagens consiste em melhorar o aspecto visual de certas feições estruturais para o analista humano e fornecer outros subsídios para a sua

interpretação, inclusive gerando produtos que possam ser posteriormente submetidos a outros processamentos. Na área de medicina, o uso de técnicas de processamento de imagens pode prover planejamento de tratamentos, simulação de cirurgias, recuperação de exames por conteúdo de imagens, dentre outras tarefas. Além disso, as principais características de um exame em formato digital podem ser extraídas, fazendo com que o computador aprenda a reconhecer anormalidades, auxiliando o médico em casos de difícil detecção.

Surgiram então os esquemas de detecção (*Computer-Aided Detection - CAD*) e diagnóstico (*CADx*) baseados em computador, que vêm ganhando cada vez mais espaço na medicina moderna. Sistemas CAD são sistemas que auxiliam a detecção de anormalidades sem, contudo, realizar qualquer tipo de diagnóstico sobre as mesmas. Tais sistemas chamam a atenção do médico para regiões que merecem a sua atenção. Por outro lado, sistemas CADx, classificam as estruturas segmentadas sugerindo o caráter benigno ou maligno da anormalidade detectada.

Neste trabalho, é proposta uma metodologia para a detecção de massas em imagens de mamografia. Convém frisar que a detecção auxiliada por computador não tem a intenção de substituir o radiologista. Entretanto, numerosos estudos indicam que, não importando o nível de habilidade, todos os radiologistas falham em detectar uma anormalidade em uma hora ou outra. Estudos recentes mostram que o índice de detecção da presença de câncer de mama poderia ser aumentado de 5 a 15% se ferramentas CAD fossem utilizadas (KOPANS, 2000, p.831). Dessa maneira, um sistema CAD funciona como um segundo leitor e aumenta o índice de detecção da presença de cânceres, podendo salvar

vidas que seriam perdidas sem um diagnóstico precoce.

Um sistema CAD geralmente divide-se em três etapas principais (Figura 1.1): segmentação da imagem de entrada para a obtenção de regiões de interesse; caracterização das regiões de interesse por meio de textura e/ou geometria; classificação das regiões de interesse através de um classificador, com o treinamento supervisionado baseado em diagnósticos feitos por um radiologista.

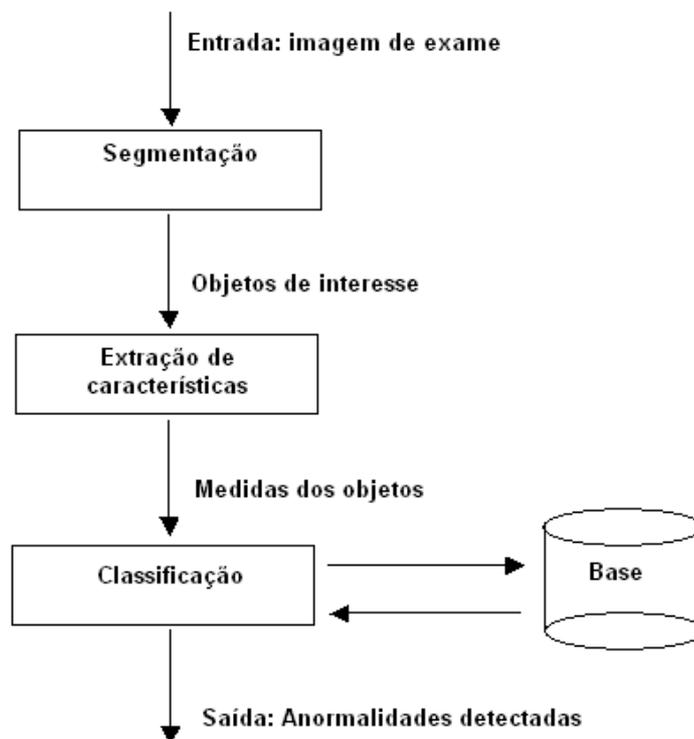


Figura 1.1 – Esquematização do funcionamento de um software CAD típico

A segmentação é usada para dividir o exame em imagens menores, de atributos similares. Em uma mamografia, por exemplo, a segmentação é usada para separar as possíveis anormalidades do restante da imagem. Tal tarefa pode ser extremamente difícil, visto que a proximidade entre tecido normal e anormal é muito grande na maioria dos casos.

Uma vez que a anormalidade em potencial está separada do restante da imagem, o sistema então extrai as principais características de textura e geometria dessa região suspeita. Na etapa seguinte, o objeto é submetido à avaliação de um classificador, que, baseado em um treinamento realizado previamente, informa ao sistema se o objeto em questão corresponde ou não a um tecido anormal.

Este trabalho tem por objetivo avaliar a eficácia do uso do algoritmo *Growing Neural Gas* (GNG) (FRITZKE, 1995) e da função *K* de Ripley (RIPLEY, 1977) na tarefa de detecção de massas benignas e malignas em imagens mamográficas. *Growing Neural Gas* é um algoritmo de agrupamento incremental e não-supervisionado, proposto por Bernd Fritzke, sendo utilizado na etapa de segmentação das estruturas da mama. A topologia do grafo resultante da aplicação desse algoritmo reflete fortemente a distribuição probabilística das amostras de entrada, razão pela qual o mesmo apresenta resultados bastante satisfatórios.

A função *K* de Ripley é uma estatística comumente utilizada em Ecologia, para descrever a distribuição espacial de árvores e outras espécies em uma floresta. Neste trabalho, investiga-se a utilização dessa função na tarefa de descrever a textura dos objetos segmentados.

Com base em experiências anteriores é constatável que o processo de detecção de anormalidades em imagens médicas é um grande desafio. Rezai-Rad (2005) apresenta uma abordagem para a detecção de microcalcificações em mamografias digitais empregando redes neurais em conjunto com a transformada de *wavelet*, com sensibilidade em torno de 94%. Belloti *et al.* (2006) apresentam uma ferramenta para a detecção de massas utilizando um algoritmo de segmentação baseado em bordas para

a seleção de regiões suspeitas. Medidas de segunda ordem obtidas através de matrizes de co-ocorrência são utilizadas para descrever a textura dessas regiões. Finalmente, redes neurais artificiais são utilizadas para a etapa de classificação das mesmas. Os resultados indicam que a ferramenta é capaz de obter uma sensibilidade de 80% na detecção de massas, com uma taxa de 4,23 falsos positivos por imagem. Chan *et.al* (2005) propõem um sistema CAD baseado na análise de mamografias tridimensionais obtidas através de um processo de tomossíntese. A área média abaixo da curva ROC ficou em torno de 0,91, com desvio padrão de 0,03. A ferramenta obteve uma sensibilidade de 85%, com 1,32 falsos positivos por imagem.

Atualmente, existem sistemas CAD aprovados pela agência americana FDA (*Food and Drugs Administration*). O software ImageChecker<sup>®</sup>, da empresa R2 Technology apresenta uma acurácia de 98,5% para calcificações e especificidade de 0,74 falsos positivos por caso. Para as massas, esse sistema apresenta uma sensibilidade de 85,7%, com uma taxa de 1,32 falsos positivos por caso. O sistema CAD MammoReader<sup>™</sup> reporta uma sensibilidade global de 89,3% (91% para casos onde calcificações são o único indício de anormalidade e 87,4% para o restante dos casos). O sistema obteve uma taxa de 1,53 verdadeiros positivos por caso e 2,32 falsos positivos entre casos com câncer e 3,32 falso-positivos entre casos sem câncer. A empresa CADx Medical Systems desenvolveu um sistema chamado de SecondLook<sup>™</sup>, o qual obteve sensibilidade de 85% para o rastreamento de cânceres (combinação de massas e microcalcificações). Adicionalmente, o sistema identificou a localização de massas malignas em 26,2% das mamografias dois anos antes do

diagnóstico de câncer (SAMPAT *et al.*, 2005).

A função  $K$  de Ripley teve sua eficácia analisada pelo autor do presente trabalho na extração de textura para a classificação de tecidos da mama em massa ou não-massa (MARTINS *et al.*, 2007a). A metodologia obteve uma taxa de acerto de 94,25%, com sensibilidade de 94,59% e especificidade de 94% em regiões de interesse segmentadas manualmente, utilizando a técnica Máquina de Vetor de Suporte na etapa de classificação. Silva *et al.* (2006) investigaram o uso dessa função na extração de características de textura em nódulos pulmonares. A etapa de classificação foi realizada através da técnica da análise discriminante. Os melhores resultados obtidos foram 97,4% de acerto, com 90% de sensibilidade e 100% de especificidade. Em Martins *et al.* (2007b) a função  $K$  de Ripley é utilizada no problema de diagnóstico de massas em mamografias digitalizadas utilizando máquinas de vetor de suporte para a classificação das mesmas em benignas ou malignas. Os melhores resultados obtidos foram 94,94% de acurácia, 92,86% de sensibilidade e 93,33% de especificidade.

A principal contribuição deste trabalho é apresentar uma metodologia para a detecção de massas em imagens mamográficas utilizando para tanto o algoritmo GNG para a segmentação e a função  $K$  de Ripley para a descrição de textura. A combinação dessas duas técnicas para esse tipo de tarefa representa uma inovação e uma contribuição para a literatura da área de sistemas CAD. Outra contribuição que merece ser citada é a adaptação de medidas geométricas originalmente propostas para três dimensões por Sousa *et al.* (2007). Neste trabalho, foram utilizadas duas das medidas geométricas propostas para a caracterização

e classificação do objeto segmentado em uma imagem mamográfica, com o objetivo de reduzir os candidatos a massa. Finalmente, este trabalho apresenta também uma modificação da função  $K$  de Ripley, proposta originalmente por Silva (2006) para a caracterização e diagnóstico de nódulos pulmonares. Tal modificação consiste em alterar a área de estudo da função  $K$  de Ripley de esferas para camadas. A contribuição deste trabalho consiste em adaptar a modificação proposta para duas dimensões, com o objetivo de extrair textura dos objetos segmentados para uma posterior classificação.

A qualidade dos resultados obtidos a partir deste trabalho poderá tornar possível a incorporação da presente metodologia em uma ferramenta para a área médica, que possa servir como uma segunda opinião ao profissional, especialmente em casos de difícil visualização e identificação da anormalidade.

Este trabalho será composto de mais quatro capítulos, descritos resumidamente a seguir.

No Capítulo 2, é exposta a fundamentação teórica necessária para a compreensão do presente trabalho. Serão descritas a extração de textura através da função  $K$  de Ripley, as principais medidas geométricas usadas para descrever as regiões de interesse e o funcionamento do algoritmo *Growing Neural Gas*, além da técnica de classificação denominada Máquina de Vetor de Suporte.

O Capítulo 3 descreve a metodologia utilizada para a detecção de massas em mamografias digitais, utilizando o algoritmo GNG, as características de textura e geometria, e finalmente, máquinas de vetor de suporte.

No Capítulo 4 são apresentados e discutidos os resultados obtidos através da metodologia proposta.

O Capítulo 5 apresenta a conclusão sobre o trabalho, mostrando a eficiência dos métodos utilizados e apresentando sugestões para trabalhos futuros.

## 2 Fundamentação Teórica

Neste capítulo são vistos os fundamentos necessários à compreensão dos métodos adotados no presente trabalho. São examinados o exame de mamografia, os tipos de anormalidades mais comuns encontradas no tecido mamário, o processamento de imagens em medicina, os métodos de análise de imagens baseados em textura, o uso da função  $K$  de Ripley para descrever textura, o funcionamento do algoritmo Growing Neural Gas e, por fim, a classificação de dados através de máquinas de vetor de suporte.

### 2.1 Mamografia

A mamografia (ou radiografia da mama) é atualmente o único exame com potencialidade comprovada para detectar o câncer clinicamente oculto da mama em tamanho e estadiamento precoces. Por isso, também é o único exame capaz de reduzir a mortalidade através do rastreamento de mulheres assintomáticas (KOPANS, 2000, p. 375).

Grande parte da estrutura da mama é formada por tecido adiposo, que é radiolucido. Isto significa que esse tipo de tecido é permeável à incidência de raios X. Por outro lado, os principais componentes da densidade radiográfica na mamografia são os tecidos conjuntivos, que são responsáveis pela maior parte das variações grosseiras de densidade. A Figura 2.1 apresenta uma imagem de mamografia.

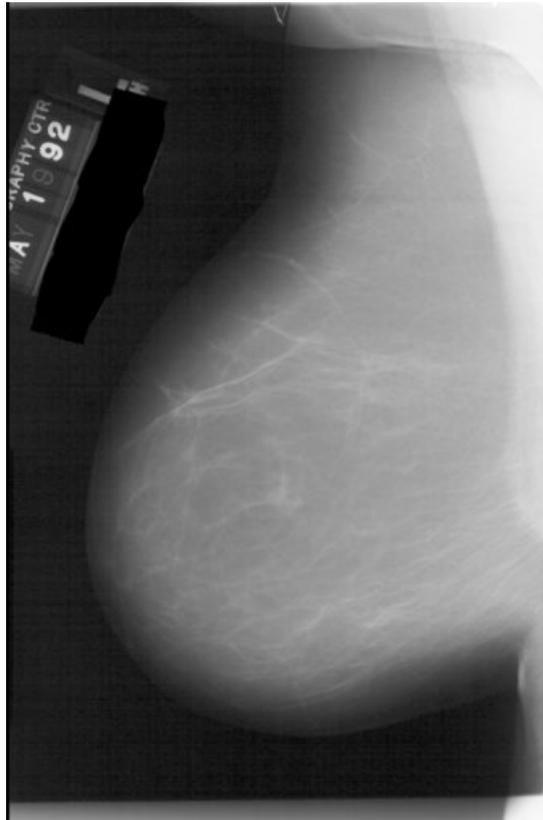


Figura 2.1 – Exemplo de mamografia. Fonte: *Digital Database for Screening Mammography - DDSM*

O objetivo da mamografia é produzir imagens de alta resolução das estruturas internas da mama, a fim de permitir a detecção do câncer de mama. Devido ao fato de que as diferenças de contraste entre tecidos doentes e normais são muito pequenas, esse exame requer um equipamento capaz de realçar tais diferenças e fornecer uma resolução de alto contraste.

Por isso, a mamografia deve ser feita em aparelho de raios X específico, chamado mamógrafo. Nele, a mama é comprimida de forma a fornecer melhores imagens e, portanto, melhor capacidade de diagnóstico.

Até que sejam planejados sistemas que possam mecanicamente posicionar a mama, são necessários técnicos altamente especializados para posicionar a paciente a fim de obter uma imagem otimizada. A compressão é necessária para evitar a subexposição da base e a superexposição dos tecidos anteriores da mama, mais finos.

Normalmente a mamografia é bilateral, ou seja, é feita uma radiografia de cada mama. Além disso, em um exame de mamografia, duas incidências ou projeções de cada mama são indispensáveis: uma visão médio-lateral oblíqua e uma crânio-caudal (Figura 2.2).

A incidência médio-lateral oblíqua é a mais útil, pois permite a visualização do alto da axila para baixo, incluindo a prega inflamamária. O termo oblíquo não se aplica à paciente, mas ao plano de compressão da mama. Se a mama não estiver bem posicionada, pode-se deixar de detectar anormalidades. O músculo peitoral deve ser visível, estendendo-se obliquamente até a metade superior da imagem. Além disso, deve ser muito largo no alto e ir se afinando à medida que cruza a parte superior da mama. Estudos sugerem que a mama é representada de maneira ótima quando o músculo peitoral é visível até o eixo do mamilo (KOPANS, 2000, p. 183).

A projeção crânio-caudal é a segunda projeção que deve ser obtida rotineiramente. O principal objetivo dessa projeção é obter uma visão da região póstero-medial da mama, complementando assim a visão médio-lateral oblíqua.

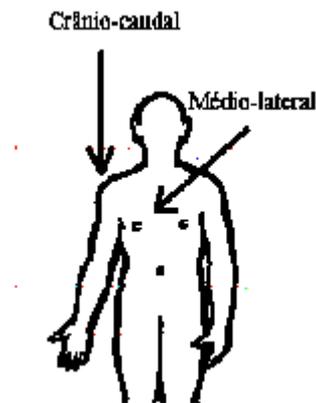


Figura 2.2 – Principais incidências de raios X em um exame de mamografia (SILVA, 2007)

A imagem dos raios X é criada a partir da absorção variável dos raios X pelos tecidos normais e anormais que resultam em números variáveis de fótons de raios X que passam através da mama e saem do outro lado. São as absorções e transmissões variáveis desses fótons pelos tecidos mamários que resultam nas sombras da imagem dos raios X.

Ao longo do tempo, foram desenvolvidas películas fluorescentes que convertiam os fótons dos raios X em luz. A luz da película era mais eficiente na exposição do filme. A eficiência ganha pelo acoplamento de uma película com o filme permitiu uma redução maior na dose de radiação requerida para representar a mama.

Dentre as vantagens do filme, pode-se dizer que é relativamente barato, e, quando acoplado com uma película, é bastante eficiente em produzir uma imagem. O filme mamográfico tem a habilidade teórica de ter 17 a 20 pares de linha por milímetro, permitindo a avaliação de detalhes muito finos. A quantidade de informação contida em uma parte do filme é enorme, tornando-o um excelente meio de armazenamento. O filme permite o armazenamento de muita informação em um espaço muito pequeno, e a informação pode ser mostrada de forma muito barata.

Entretanto, há diversas desvantagens nas combinações de filme/película. Em consequência do fato de que o filme é o detector e o meio visualizador, as exigências reais da exposição são ditadas pela necessidade de formar uma imagem visível. O filme não responde linearmente aos fótons incidentes. Não é muito sensível aos fótons precoces que incidem, de modo que as áreas da mama de absorção elevada de raios X (transmissão baixa de fótons ao filme) todas parecem similarmente brancas na imagem. É somente sobre uma escala razoavelmente estreita da exposição que alterações pequenas na atenuação resultam em diferenças no escurecimento do filme (contraste). Uma vez que as estruturas normais da mama e dos cânceres de mama têm atenuação similar aos raios X, esta limitação pode dificultar a visualização e a separação de estruturas normais e anormais. O contraste é extremamente importante. Embora o filme tenha resolução espacial elevada, se houver um contraste insuficiente entre duas estruturas pequenas, então elas não serão visíveis separadamente.

Uma outra desvantagem do filme é que uma vez que a imagem é revelada, torna-se imutável. O filme é o único registro da imagem, e se for perdido ou danificado, então a perda é irrecuperável. O armazenamento de filmes de raio X requer quantidades de espaço grandes, e os estudos são freqüentemente perdidos. Se for desejada uma consulta com um outro médico ou perito, as imagens devem ser fisicamente fornecidas ao consultante.

A imagem produzida através da mamografia convencional é analógica, ou seja, os componentes de uma imagem em filme são razoavelmente contínuos. A mamografia digital produz uma imagem digital,

que pode ser imaginada como a amostragem de um sinal bidimensional contínuo em pontos discretos, formando uma grade de linhas e colunas. Cada ponto da grade é chamado de *pixel* (*picture element*). Quanto mais próxima a amostragem, mais nítida a imagem; quanto mais grosseira a amostragem, pior é a qualidade da imagem.

Dois abordagens estão sendo utilizadas na obtenção de imagens de mamografias digitais: a digitalização de mamografias convencionais de filme/película (utilizada para a realização deste trabalho) e a aquisição direta (ou primária) de dados digitais da mama por detectores em estado sólido.

Os digitalizadores trabalham passando a luz através da radiografia e gravando a intensidade da luz transmitida em cada ponto da fotografia. Em seguida, um número é atribuído a essa intensidade luminosa. Depois de copiada para a forma digital, a imagem pode ter seu comprimento e largura variados, sendo possível manipular por computador os dados adquiridos através do sistema de filme/película.

Porém, existem também algumas desvantagens na digitalização do filme. Apesar de permitir o realce da visibilidade das estruturas da mama, a digitalização de mamografias obtidas através do sistema filme/película não adiciona qualquer informação além daquelas que já estavam presentes na radiografia. Além disso, a imagem digitalizada tem menos informação do que a imagem original uma vez que algumas informações sutis são perdidas nesse processo.

Por outro lado, a aquisição digital direta (Figura 2.3) apresenta um enorme potencial em oferecer uma maior profundidade de informação. Uma vez que é possível subtrair tecido mamário e melhorar a clareza das

calcificações, a detecção precoce do câncer de mama será com certeza realçada. O desafio está em determinar a melhor maneira em apresentar a informação ao observador, uma vez que as possibilidades serão praticamente ilimitadas.

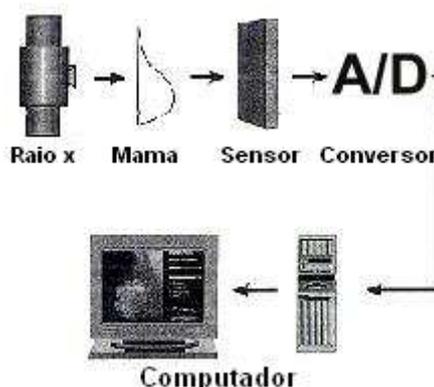


Figura 2.3 – Diagrama de um mamógrafo digital (SILVA, 2007).

Os maiores impedimentos para a disseminação da mamografia digital encontram-se nos altos custos dos equipamentos, o que acaba se refletindo no custo do exame para a paciente, e na grande quantidade de memória necessária para armazenar as imagens. Além disso, os dispositivos de visualização (monitores) de baixo custo não são capazes de exibir uma mamografia digital inteira, o que eleva ainda mais o custo total do equipamento.

### 2.1.1 Massas e outras anormalidades

Os vários tipos de anormalidades observáveis através da mamografia incluem massas ou lesões, distorções de arquitetura, e calcificações (Figura 2.4). As massas, que representam o tumor em si, aparecem como regiões densas, de tamanho e formato variáveis. Podem ser classificadas, de acordo com o aspecto de suas bordas, como

circunscritas, microlobuladas, obscurecidas, mal definidas e espiculadas (Figura 2.5). Com relação ao formato, podem ser classificadas em redondas, ovais, lobulares ou irregulares. (Figura 2.6).

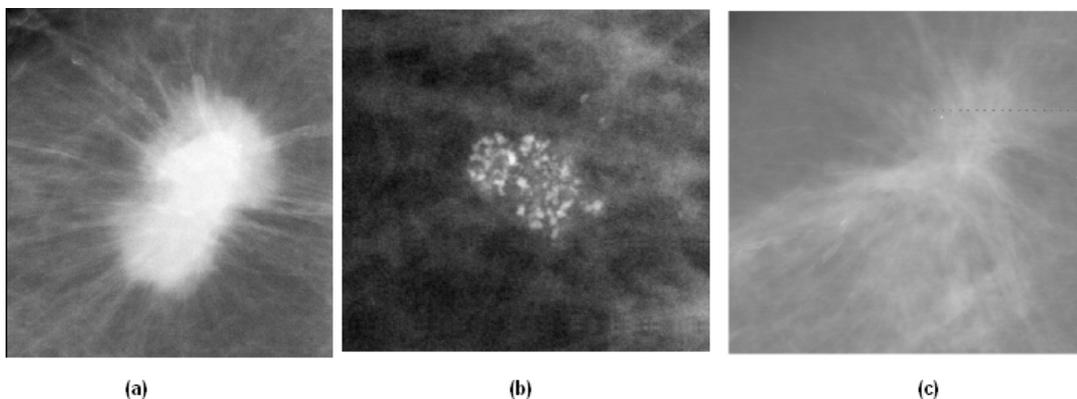


Figura 2.4 – Anormalidades encontradas em tecidos mamários. (a) massa espiculada; (b) agrupamento de microcalcificações e (c) distorção de arquitetura. Fonte: *Digital Database for Screening Mammography – DDSM*

As massas benignas são geralmente bem definidas e com pouca repercussão na arquitetura vizinha da mama. Massas que possuem fronteiras irregulares, e que se confundem com tecidos adjacentes têm grandes chances de terem caráter maligno (VALLE, 1999). Entretanto, o diagnóstico de massas através da mamografia não é tão simples na maioria dos casos, a menos que as massas apresentem sinais característicos de um processo maligno, ou, no caso das massas benignas, calcificações benignas e/ou bordas encapsuladas. Por essa razão, a mamografia é considerada uma excelente técnica de detecção (rastreamento), porém não é tão boa para a realização de diagnósticos (definição de caráter benigno ou maligno de uma massa) (KOPANS, 2000, p. 211).

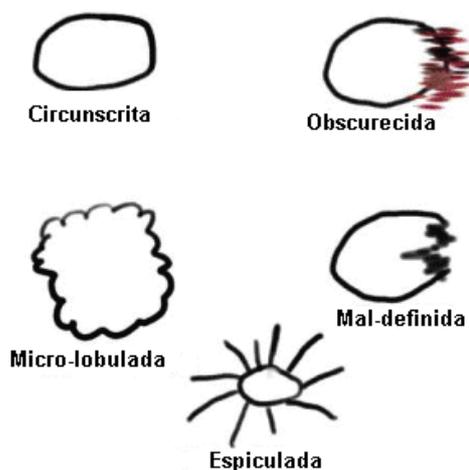


Figura 2.5 – Classificação das massas de acordo com o aspecto de suas bordas. Adaptado de Padwal (2007).

O tumor de mama nem sempre produz uma massa mamograficamente visível. Ele pode produzir somente uma área de distorção de arquitetura. De modo geral, o fluxo das estruturas na mama é dirigido para o mamilo. Os distúrbios nesse fluxo devem ser avaliados com cuidado, embora possam ser causados por fatores benignos (KOPANS, 2000, p. 299).

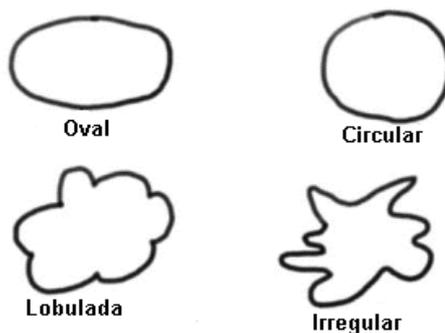


Figura 2.6 – Classificação das massas de acordo com sua forma. Adaptado de Padwal (2007).

Em conjunto com os tumores benignos e malignos da mama,

freqüentemente estão associadas calcificações ou microcalcificações, que são pequenos depósitos de cálcio. As calcificações aparecem como pontos brilhantes nas mamografias e podem variar em forma e tamanho. As formas das partículas e a heterogeneidade das formas e dos tamanhos são valiosas pistas na determinação da causa provável dos depósitos. Calcificações benignas geralmente são maiores, regulares e esparsas. Já as malignas são irregulares no tamanho, agrupadas e localizadas em extensa área. As causas das calcificações malignas são a secreção celular ou a calcificação de células cancerígenas. Dentre as causas da calcificação benigna estão a arterioesclerose, a necrose de gordura, o fibroadenoma involuído e mastite com deposição de cálcio (VALLE, 1999).

## **2.2 Processamento de imagens digitais**

O processamento de imagens digitais é compreendido como um conjunto de técnicas computacionais que englobam desde a aquisição da imagem digital até seu reconhecimento e interpretação por parte de uma máquina digital. Nesse contexto, uma imagem digital é definida como uma função bidimensional  $f(x,y)$ , sendo  $x$  e  $y$  as coordenadas espaciais e a amplitude de  $f$  para um dado ponto  $(x,y)$  é chamada de intensidade ou *pixel* (*picture element*). É importante ressaltar que todos os valores de  $x$ ,  $y$  e respectivas intensidades devem ser quantidades finitas e discretas, e que uma imagem digital apresenta um número finito de pontos  $x$  e  $y$  (GONZALEZ, 1992).

O interesse em métodos de processamento imagens digitais surgiu, principalmente, da necessidade de melhorar a qualidade da

informação pictorial para interpretação humana. Ao longo das duas últimas décadas, a área de processamento de imagens digitais apresentou um rápido crescimento. A evolução da tecnologia de computação digital, bem como o desenvolvimento de novos algoritmos para lidar com sinais bidimensionais estão permitindo uma gama de aplicações cada vez maior.

Através das técnicas de processamento de imagens digitais é possível melhorar o aspecto visual de certas feições estruturais para o observador humano e fornecer outros elementos para a interpretação visual da imagem, podendo inclusive gerar outros produtos que possam ser posteriormente submetidos a outros processamentos. Além disso, o reconhecimento e interpretação por parte da máquina propiciam inúmeras vantagens. A visão humana é sensível somente a uma pequena faixa do espectro eletromagnético, enquanto que máquinas podem ser projetadas para cobrir desde radiação gama até ondas de rádio. Programas computacionais podem ser treinados para interpretar imagens que são apresentadas difícil interpretação por parte de seres humanos, tais como as imagens de ultra-som e microscopia eletrônica (GONZALEZ, 1992). Tudo isso faz com que a área de processamentos de imagens digitais englobe um largo e variado campo de aplicações. Exemplos de áreas em que tais técnicas computacionais têm sido empregadas com sucesso são: astronomia, medicina, física, arqueologia, biologia, defesa e aplicações industriais.

Na medicina, por exemplo, o processamento de imagens digitais permite a coloração ou a intensificação de contraste para facilitar a interpretação de radiografias e outras imagens biomédicas e até mesmo realizar detecções e diagnósticos de maneira automática, fornecendo

assim uma segunda opinião aos profissionais de saúde.

### **2.2.1 Passos fundamentais do processamento de imagens digitais**

O processamento de imagens digitais se inicia com a aquisição da imagem digital. Para tanto, é usado um sensor para imageamento e um conversor analógico/digital para digitalizar o sinal produzido pelo sensor.

O próximo passo trata de pré-processar aquela imagem. O principal objetivo do pré-processamento é melhorar a imagem de forma a aumentar as chances para o sucesso dos processos seguintes. O pré-processamento tipicamente envolve técnicas para o realce de contrastes e remoção de ruído.

O estágio seguinte trata da segmentação, a qual pode ser definida como um processo que 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. Neste trabalho, foi utilizada a segmentação manual das regiões de interesse da mama.

O processo de representação e descrição, também chamado seleção de características, procura extrair características que resultem em alguma informação quantitativa de interesse ou que sejam básicas para discriminação entre classes de objetos. O último estágio envolve reconhecimento e interpretação. Reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo seu descritor. A interpretação envolve a atribuição de significado a um conjunto de objetos reconhecidos. A Figura 2.7 mostra o relacionamento entre os passos fundamentais em processamento de imagens digitais.

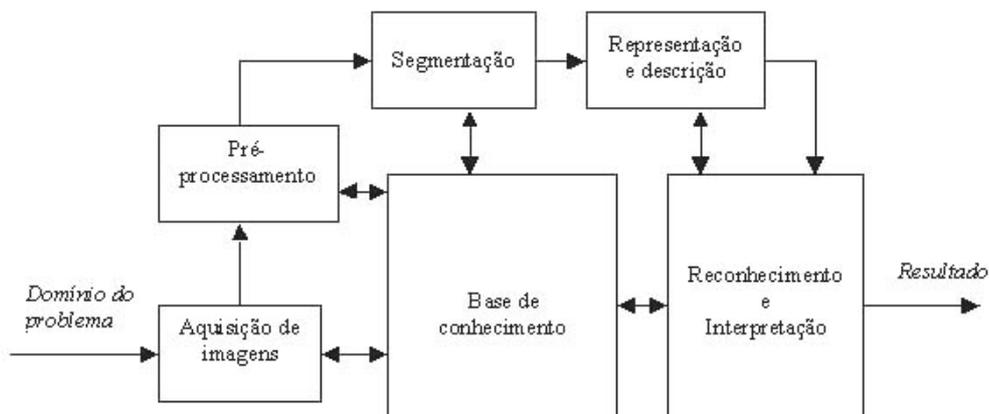


Figura 2.7 – Passos fundamentais em processamento de imagens digitais.

Adaptado de (GONZALEZ, 1992).

Neste trabalho foram utilizados os seguintes passos: segmentação, representação e descrição, reconhecimento e interpretação.

## 2.2.2 Equalização de histograma

Equalizar o histograma significa, na prática, maximizar a variância do histograma de uma imagem, obtendo assim um melhor contraste. Por contraste, entende-se uma medida qualitativa relacionada à distribuição dos tons de cinza em uma imagem.

Para uma imagem em tons de cinza, o histograma é uma função discreta  $h(L)$ , a qual retorna o número de ocorrências de cada nível de cinza  $0 \leq L \leq 2^b - 1$  na imagem, sendo  $b$  o número máximo de *bits* capaz de representar cada *pixel*. O histograma é comumente normalizado no intervalo  $[0,1]$  dividindo-se  $h(L)$  pelo número  $N \times M$ , sendo  $N$  e  $M$  as dimensões da imagem. Nesse sentido, o histograma normalizado

representa a distribuição de probabilidade dos valores dos *pixels*, pois a soma de todas as componentes de um histograma normalizado é igual a 1.

O histograma acumulado (Figura 2.8) de uma imagem é uma função  $h_a(L)$  que produz o número de ocorrências de níveis de cinza menores ou iguais a  $L$ . Assim, o histograma acumulado fornece não apenas o número de ocorrências do nível desejado, mas a soma de todas as ocorrências dos níveis inferiores ao nível desejado, acrescida do número de ocorrências do próprio. O histograma acumulado pode ser obtido diretamente do histograma da imagem. Dessa maneira, utilizando o histograma normalizado, obtém-se o correspondente histograma acumulado normalizado.

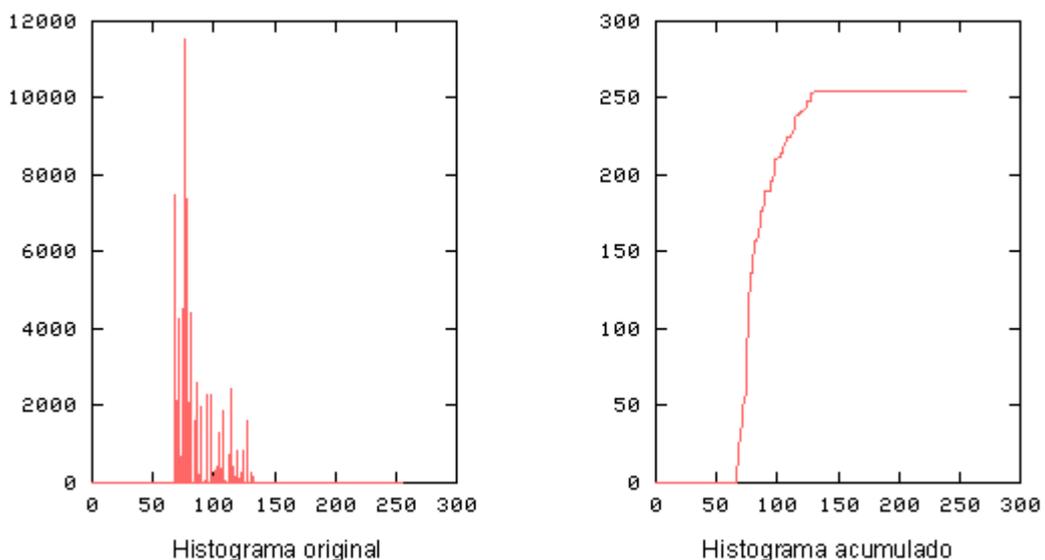


Figura 2.8 – Representação de um histograma e seu respectivo histograma acumulado

A equalização do histograma (Figura 2.9) em um imagem tem por objetivo realçar a imagem, obtendo uma aproximação de um histograma uniforme. Existem dois tipos básicos de equalização (global ou

local) (GONZALEZ, 1992). Neste trabalho, a equalização global do histograma é utilizada para melhorar o contraste dos objetos segmentadas, fornecendo assim uma melhor descrição da textura de cada objeto.

Para a realização desse processo, é utilizado o histograma acumulado normalizado. O nível de cinza  $L_{eq}$  de um pixel  $p$  na imagem equalizada é dado por:

$$L_{eq}(p) = h_{anorm}(L(p)) * (2^b - 1) \quad (2.1)$$

sendo  $L_{eq}(p)$  o nível de cinza do pixel  $p$  na imagem original,  $h_{anorm}(L(p))$  o histograma acumulado normalizado do nível de cinza  $L(p)$  e  $b$  é o número de *bits* utilizado para representar cada *pixel* na imagem original.

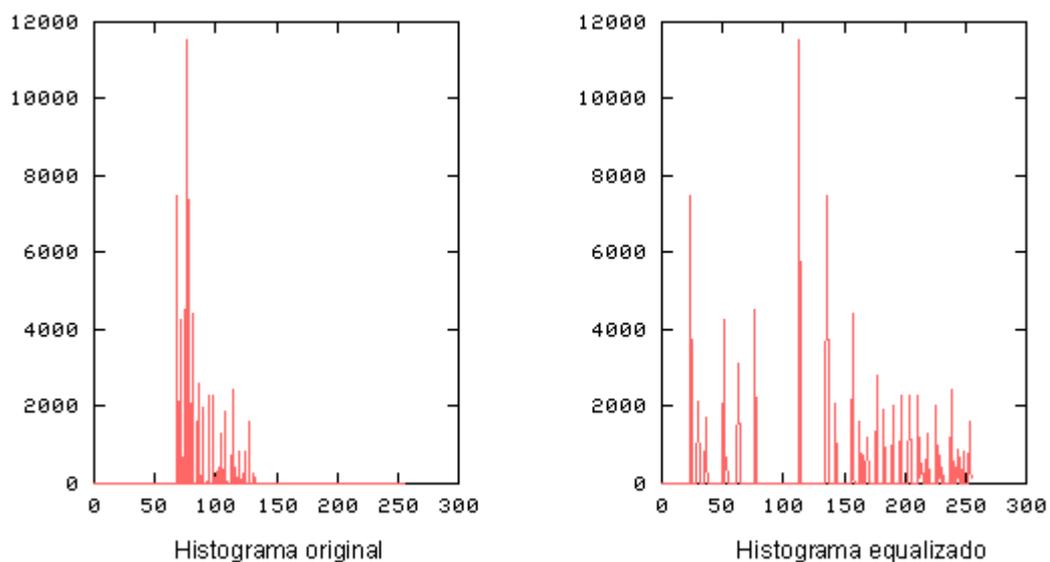


Figura 2.9 – Representação de um histograma e seu respectivo histograma equalizado

### 2.2.3 – Limiarização (*Thresholding*)

O processo de limiarização ou *thresholding* pode ser considerado uma forma simples de segmentação, a qual consiste em particionar uma imagem em regiões com características semelhantes.

Existem basicamente dois tipos de limiarização: a limiarização global e a limiarização adaptativa. Dado um valor de limiar  $t$  e o número de bits por *pixel*  $b$ , a limiarização global divide a imagem em duas regiões, uma com os níveis de cinza entre 0 e  $t$ , e outra com níveis de cinza entre  $t+1$  e  $2^b-1$ . A limiarização adaptativa, por outro lado, busca encontrar limiares diferentes para regiões diferentes da imagem, melhorando a qualidade do resultado final. Diversas técnicas podem ser utilizadas para encontrar o melhor valor do limiar para uma imagem (GONZALEZ, 1992). Uma das mais simples consiste em examinar o histograma da imagem, verificando se existem “vales” ou “depressões” entre dois máximos. O valor mínimo entre esses dois máximos é um bom candidato a limiar global.

Dessa maneira, a imagem limiarizada é uma imagem binária, onde os pixels com níveis de cinza entre 0 e  $t$  recebem o valor 0, e o restante dos pixels recebem o valor 1. Normalmente, são utilizados os valores 0 e 255 para facilitar a visualização do resultado em imagens de níveis de cinza.

Neste trabalho, utilizou-se um processo de limiarização global para separar a mama propriamente dita dos elementos da mamografia que não pertencem à mama, tais como os rótulos que exibem informações sobre o exame.

## **2.3 Análise de imagens baseada em textura**

A característica de textura é uma fonte importante de informações para o processo de análise e interpretação de imagens. Os métodos de análise de imagens baseados em características de textura têm sido amplamente utilizados no meio científico e industrial, possuindo aplicações na medicina, microscopia, sensoriamento remoto, controle de qualidade, recuperação de dados em bases gráficas, dentre outras.

### **2.3.1 Textura**

Apesar de não existir uma definição formal, textura é visualmente compreendida como uniformidade, densidade, aspereza, regularidade, intensidade, dentre outras características da imagem. Analistas são capazes de extrair visualmente informações de textura de uma imagem, mas não é fácil para eles estabelecer modelos para descrever matematicamente esse conceito intuitivo.

A textura é compreendida como um mapeamento entre as propriedades primitivas das tonalidades e os relacionamentos espaciais entre elas. Os conceitos de tonalidade e textura não são independentes, sendo que em algumas imagens a tonalidade é dominante e em outras a textura é dominante.

As três abordagens principais usadas em processamento de imagens para a descrição de texturas são a estrutural, a espectral e a estatística (GONZALEZ, 1992). Neste trabalho foi utilizada a abordagem estatística.

A abordagem estrutural utiliza a idéia de que texturas são compostas de primitivas dispostas de forma aproximadamente regular e repetitiva, de acordo com regras bem definidas. Como exemplo, a descrição da textura baseada em linhas paralelas regularmente espaçadas;

Por outro lado, a abordagem espectral baseia-se em propriedades do espectro de Fourier, sendo principalmente utilizada na detecção de periodicidade global em uma imagem através da identificação de picos de alta energia no espectro.

Na abordagem estatística, a textura é definida por um conjunto de medidas locais extraídas do padrão. A análise de imagens e extração de suas características através da abordagem estatística propicia uma descrição da mesma através de regras estatísticas que governam a distribuição e a relação entre seus níveis de cinza. Essa foi a abordagem utilizada para uso neste trabalho, porque as imagens de tecidos das mamas podem ser tratadas como texturas não repetitivas. Medidas estatísticas comuns incluem contraste, energia, entropia, correlação, homogeneidade, momento, variância, obtidas através de Matrizes de Co-Ocorrência (HARALICK, 1973).

A distribuição dos níveis de cinza dos *pixels* pode ser descrita por estatísticas de primeira ordem, como média, variância, desvio padrão ou estatísticas de segunda ordem, como a probabilidade de dois *pixels* terem um determinado nível de cinza ocorrendo com um relacionamento espacial particular. Neste trabalho, a análise de textura é feita através da relação espacial entre pontos (*pixels*) individuais quantificada através da função  $K$  de Ripley, a qual será discutida na próxima seção.

### 2.3.2 Função $K$ de Ripley

A maioria dos sistemas do mundo real não são espacialmente homogêneos, mas apresentam um certo tipo de correlação espacial. A análise de padrões de pontos (URBAN, 2003) engloba um certo número de ferramentas para examinar detalhadamente a distribuição de pontos discretos, como por exemplo, *pixels* em uma imagem mapeados para coordenadas cartesianas  $(x,y)$  em uma região de interesse.

A análise desses dados espaciais envolve a exploração de padrões e relacionamentos entre os mesmos, além da busca de explicações para estes padrões e relacionamentos. O objetivo na análise de padrões de pontos espaciais é de verificar se existe uma tendência nos eventos observados de exibirem um padrão sistemático, como regularidade, agrupamentos ou aleatoriedade (PAIVA *et al.*, 1999).

Nesse contexto, um dos métodos de análise de segunda ordem comumente utilizados é a função  $K$  de Ripley. Essa função pode ser utilizada para resumir um padrão de pontos, testar hipóteses sobre o padrão, estimar parâmetros e ajustar modelos (RIPLEY, 1977).

Dado um raio  $t$ , a função  $K$  é definida como

$$K(t) = \frac{A}{n^2} \sum_i \sum_j \delta_{ij}(d_{ij}) \quad (2.2)$$

para,  $i \neq j$ , sendo  $A$  a área amostrada,  $n$  o total de pontos e  $\delta_{ij}$  uma função que devolve 1 se a distância  $d_{ij}$  entre os pontos  $i$  e  $j$  é menor do que  $t$  e 0 em caso contrário. Em outras palavras, a função  $K$  de Ripley conta o número de ocorrências do evento  $j$  em um círculo de raio  $t$  para cada centro  $i$ , conforme visto na Figura 2.10.

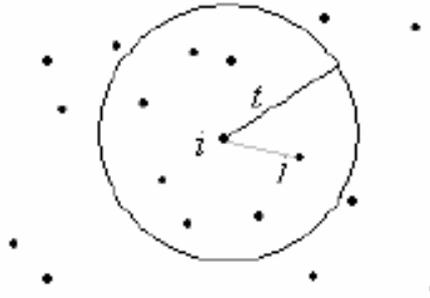


Figura 2.10 – Ilustração do funcionamento da Função  $K$  de Ripley para um dado raio  $t$

A função  $K(t)$  assume estacionariedade sobre pequenas escalas na região de interesse. O conceito de estacionariedade significa que a média e a variância do processo são independentes da localização e constante na região. Também é comum assumir isotropia, isto é, uma unidade de distância na direção vertical tem o mesmo peso que uma unidade de distância na horizontal.

Como cada ponto da região é tomado uma vez para ser o centro do círculo, a função  $K(t)$  provê uma inferência em nível global sobre a área em estudo. Entretanto, esta medida também pode ser considerada em uma forma local para o  $i$ -ésimo ponto

$$K_i(t) = \frac{A}{n} \sum_{i \neq j} \delta(d_{ij}) \quad (2.3)$$

Dessa maneira, é possível descrever a textura de uma região em uma imagem através da função local  $K$  de Ripley. A partir da escolha

de um centro  $i$ , são examinados as ocorrências de pixels de um mesmo nível de cinza  $j$ , para diferentes valores de raios  $t$ . É importante ressaltar que cada nível de cinza ou padrão espacial pontual é examinado separadamente dos demais, e tratado como a ocorrência ou não de um evento dentro da distância  $t$  especificada. Assim, o número de elementos do vetor de características obtido através do uso de  $K(t)$  é dado pelo número de níveis de cinza presentes na imagem vezes o número de raios desejado.

O presente trabalho propõe uma modificação sobre a função  $K(t)$ , através da análise dos padrões de pontos em anéis, ao invés de círculos, conforme proposto pelo autor em (MARTINS, 2007a). Basicamente, a modificação consiste em substituir a região de interesse da Equação 2.3 pela região compreendida entre dois círculos concêntricos (Figura 2.11). Tal modificação mostrou-se superior à função  $K(t)$  tradicional na caracterização de tecidos da mama para a classificação em massa e não-massa. Os melhores resultados obtidos com a função  $K$  de Ripley tradicional foram 88,51% de acerto, 90,28% de sensibilidade e 87,25% de especificidade. A modificação proposta obteve uma taxa de acerto de 94,25% de acerto, 94,59% de sensibilidade e 94% de especificidade.

A superioridade da modificação proposta pode ser explicada pelo fato de que o uso de círculos como regiões de estudo acaba fornecendo informação cumulativa para regiões periféricas do objeto. Em contrapartida, o uso de anéis concêntricos elimina possíveis interferências de regiões centrais, fornecendo informação mais precisa a respeito das regiões periféricas do objeto. Por essa razão, neste trabalho foi utilizada a modificação proposta, com a finalidade de caracterizar a textura dos

objetos segmentados pelo algoritmo *Growing Neural Gas*.

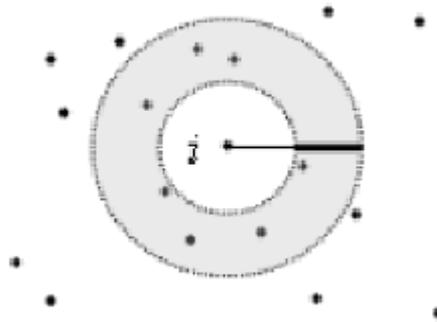


Figura 2.11 - Ilustração do funcionamento da Função  $K$  de Ripley modificada para um dado raio  $t$

## 2.4 Medidas Geométricas

Muitas vezes, em processamento de imagens, a informação de textura não é suficiente para separar adequadamente os objetos de interesse em classes distintas. Por isso, medidas geométricas são utilizadas como informação adicional para caracterizar os objetos segmentados.

Na literatura de processamento de imagens pode ser encontrada uma grande variedade de descritores ou medidas geométricas (GUELPH, 2001). Entre eles, podem ser destacados compacidade, excentricidade, circularidade, descritores Fourier, descritores baseados em momentos, com principal representante os sete momentos invariantes (*Hu Moments*), e curvatura (RANGAYYAN, 1997).

Neste trabalho, foram utilizadas as seguintes medidas geométricas para caracterizar massas e regiões normais: excentricidade, circularidade, compacidade, densidade circular e desproporção circular. Todas essas medidas são obtidas através dos objetos segmentados a

partir do algoritmo *Growing Neural Gas*. Entretanto, para o cálculo de tais medidas não são levados em conta os níveis de cinza presentes nos objetos, ou seja, o objeto é binarizado. Somente suas propriedades geométricas, tais como área e perímetro são utilizados no cálculo das medidas apresentadas a seguir.

#### 2.4.1. Excentricidade (E)

Excentricidade é definida como a razão entre o menor e maior eixo horizontal ou vertical, caracterizando como o objeto está distribuído espacialmente entre seus eixos (BRAZ JÚNIOR, 2006). Dessa maneira, quanto mais baixo o valor da excentricidade, mais circular é o objeto.

Valores demasiadamente altos indicam uma grande diferença espacial entre o maior e o menor eixo que compõe o objeto. A excentricidade pode ser calculada por

$$E = \frac{(\mu_{02} - \mu_{20}) + 4\mu_{11}}{A} \quad (2.4)$$

sendo A a área do objeto. Os momentos centrais  $\mu$  são obtidos através de

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^p (y - \bar{y})^q \quad (2.5)$$

com  $p+q > 1$  e  $(\bar{x}, \bar{y})$  sendo o centro de gravidade do objeto em estudo. A Figura 2.12 compara dois objetos, um com alta e outro com baixa excentricidade.

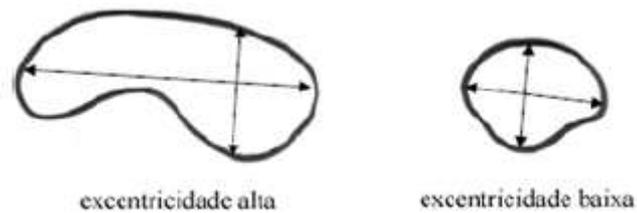


Figura 2.12 - Ilustração da medida geométrica de excentricidade. Fonte:  
(BRAZ JUNIOR, 2006)

#### 2.4.2. Circularidade (C)

É a medida geométrica que define o quão circular o objeto se apresenta. Trata-se da razão entre a área do objeto e o perímetro convexo, onde o perímetro convexo é calculado sobre a região que engloba o objeto de forma a não deixar picos ou defeitos no contorno. A circularidade é definida pela relação

$$C = \frac{4\pi A}{(p_{convexo})^2} \quad (2.6)$$

sendo  $A$  a área do objeto em estudo e  $p_{convexo}$  é o perímetro convexo. A circularidade terá valor máximo 1 para um círculo. Logo, quanto mais próximo de um círculo é a figura, mais próximo de 1 é o valor de sua circularidade (BRAZ JÚNIOR, 2006) . A Figura 2.13 exemplifica o conceito de circularidade através da indicação dos valores dessa medida para dois objetos.

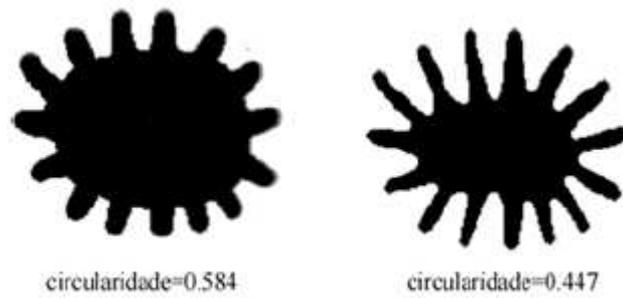


Figura 2.13 - Ilustração da medida geométrica de circularidade. Fonte: Braz Junior (2006)

### 2.4.3. Compacidade ( $C_o$ )

Essa medida mostra o quão denso o objeto é, em comparação com uma figura perfeitamente densa - o círculo (SCHOUTEN, 2003).

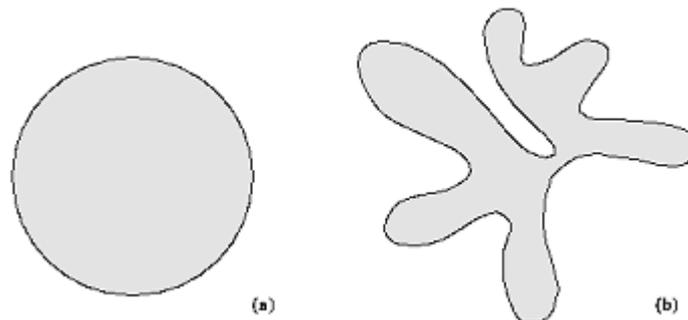


Figura 2.14 – Ilustração da compacidade de dois objetos: (a) Compacto; (b) não-compacto. Fonte: (Schouten, 2003).

A compacidade é definida por

$$C_o = \frac{p^2}{4\pi A} \quad (2.7)$$

sendo  $A$  a área do objeto em estudo e  $p$  o seu perímetro. A Figura 2.14 ilustra graficamente o conceito de compacidade.

#### 2.4.4. Desproporção Circular (D)

A análise desta medida, proposta por Sousa *et al.* (2007) para a caracterização de nódulos pulmonares em três dimensões, pode informar o quanto determinado objeto é desproporcional em relação a uma superfície totalmente circular.

A desproporção circular pode ser obtida através da seguinte equação

$$D = \frac{p}{2\pi R_e} \quad (2.8)$$

sendo  $p$  o perímetro do objeto em estudo e  $R_e$  o raio estimado do círculo de mesma área do objeto em estudo. O raio estimado  $R_e$  pode ser obtido através de

$$R_e = \sqrt{\frac{A}{\pi}} \quad (2.9)$$

sendo  $A$  a área do objeto em estudo.

#### 2.4.5. Densidade Circular (D<sub>c</sub>)

Uma medida geométrica bastante comum consiste em comparar a área de um objeto com sua *bounding box*, isto é, o menor retângulo capaz de armazenar o objeto. Tais medidas, entretanto, enfrentam o problema de que podemos ter diferentes valores para o mesmo objeto com rotação diferente, ou seja, não são rotacionalmente invariantes.

A densidade circular, por outro lado, utiliza um círculo, figura rotacionalmente invariante, para estimar qual a porcentagem do círculo que

também corresponde ao objeto (SOUSA *et al.*, 2007). Isso é feito utilizando um círculo com a mesma área do objeto e com o centro em seu centro de massa. A Figura 2.15 ilustra o conceito de densidade circular para diferentes objetos. Tal medida pode ser calculada através da seguinte fórmula:

$$D_c = \frac{100n}{A} \quad (2.10)$$

sendo  $A$  a área do objeto e  $n$  o total de pontos pertencentes ao objeto e também ao círculo de raio estimado  $R$  (Equação 2.9), com centro no centro de massa do objeto.

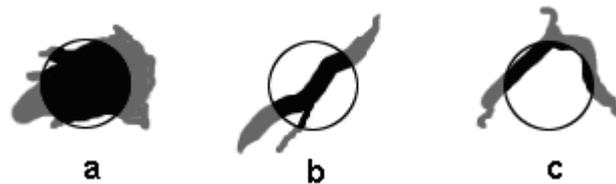


Figura 2.15 - Ilustração da densidade circular de três objetos

A densidade circular tende a assumir valores próximos a zero para objetos muito alongados e valores próximos a 100 para objetos mais arredondados.

## 2.5 Mapas auto-organizáveis

Técnicas de aprendizado não-supervisionado são bastante úteis quando deseja-se extrair informações relevantes de dados não-rotulados. Tais técnicas são capazes de extrair padrões presentes nos dados de entrada sem a necessidade de rotular previamente as amostras com classes pré-definidas.

Os mapas auto-organizáveis (*Self-Organized Maps* - SOM) constituem uma importante ferramenta para capturar características importantes contidas em um espaço de entrada não-rotulado. São caracterizados pela formação de uma mapa topográfico no qual as coordenadas dos neurônios (ou nós) são formadas pelas mesmas características contidas nos padrões de entrada. Após a aprendizagem, cada neurônio passa a responder por um subconjunto do espaço inicial de dados. Dessa maneira, os mapas auto-organizáveis são capazes de representar um conjunto grande de vetores de entrada  $x_i \in X$  em um conjunto menor de vetores  $w_j \in W$ , de modo a fornecer uma boa aproximação para o espaço de entrada original  $X$  (HAYKIN, 2001).

Outra característica importante dos mapas auto-organizáveis é que a formação topológica dos padrões de entrada é preservada. Isso significa que padrões próximos no conjunto amostral são relacionados a neurônios próximos nos vetores de saída.

De maneira geral, o processo de aprendizagem em um mapa auto-organizável consiste de três fases. Na primeira fase, competitiva, cada padrão de entrada é apresentado a todos os neurônios para que aquele mais próximo do padrão apresentado seja o vencedor (*Best Matching Unit* - BMU). Na segunda fase, cooperativa, é definida a vizinhança relativa ao neurônio vencedor. Na terceira fase, adaptativa, os vetores de código do neurônio vencedor e dos seus vizinhos serão alterados segundo algum critério de atualização.

Após a aprendizagem os vetores de código corresponderam a uma aproximação não-linear dos padrões de entrada. Existem vários algoritmos de aprendizagem que podem ser usados em conjunto com

mapas auto-organizáveis.

Uma possível aplicação dos mapas auto-organizáveis é a agrupamento ou clusterização, que consiste em separar uma massa de dados inicial em um número desejado de grupos, com a restrição de que os componentes de cada grupo compartilhem de características semelhantes. Em resumo, o agrupamento pode ser descrito como o processo de organizar uma coleção de vetores  $k$ -dimensionais em um número  $n$  de grupos cujos membros compartilham características similares segundo alguma função de medida. Depois de formados, cada um desses  $n$  grupos é representado por somente um vetor  $k$ -dimensional denominado vetor de código. Algoritmos de agrupamento podem ser usados em quantização vetorial, que é o processo de “quantizar” vetores de entrada  $k$ -dimensionais em vetores de saída também  $k$ -dimensionais.

Um exemplo de aplicação dos algoritmos de agrupamento em processamento de imagens é a segmentação, a qual consiste em dividir uma imagem em  $n$  regiões distintas de acordo com algum critério de semelhança entre essas regiões. Neste trabalho, utilizou-se o algoritmo *Growing Neural Gas* (GNG) para a segmentação da imagem de mama em objetos distintos, os quais serão classificados em massa ou não-massa através de uma máquina de vetor de suporte. O algoritmo GNG é considerado um melhoramento dos diversos mapas auto-organizáveis existentes, e será descrito detalhadamente na próxima Seção.

### **2.5.1 Algoritmo *Growing Neural Gas***

O algoritmo *Growing Neural Gas* (GNG) (FRITZKE, 1995) é

considerado uma otimização de métodos anteriores, tais como os Mapas Auto-Organizáveis de Kohonen (HAYKIN, 2001) e o algoritmo *Neural Gas* (MARTINEZ e SCHULTEN, 1991). Trata-se de um algoritmo de agrupamento que funciona de forma incremental, ou seja, o número de grupos vai aumentando a medida em que o algoritmo é executado, e não supervisionada, no sentido de que os *clusters* são formados com base apenas nas características estatísticas dos dados de entrada.

Inicialmente, o algoritmo recebe um conjunto de vetores de entrada, de dimensionalidade conhecida. Partindo de dois nós iniciais, o algoritmo passa a acrescentar novos nós a medida em que é executado, seguindo um conjunto de regras que serão descritas a seguir, e com isso cria um grafo no mesmo espaço vetorial das amostras de entrada. Cada um dos nós é um representante do *cluster*, possuindo uma posição nesse espaço dimensional das amostras. O GNG pode ser utilizado em processos de quantização, bastando para isso interpretar cada representante como um dos vetores de código.

A topologia do grafo encontrado, após inseridos todos os nós, reflete fortemente a distribuição probabilística das amostras de entrada. A cada iteração do algoritmo, uma amostra  $a_m$  pertencente à massa de dados de entrada é escolhida segundo a distribuição de probabilidades utilizada no treinamento. Essa amostra também é chamada de sinal de entrada. O neurônio mais próximo à essa amostra é então escolhido (fase de competição). Em seguida, determina-se o segundo neurônio mais próximo, criando uma ligação entre esse neurônio e o vencedor (fase de cooperação). A seguir, o vencedor e seus vizinhos são movidos em direção à amostra de entrada de acordo com parâmetros pré-definidos (fase de

adaptação).

O grafo final obtido após a execução do GNG é considerado uma triangulação induzida de Delaunay. Da mesma forma, o conjunto das regiões representadas por cada neurônio é considerado um diagrama de Voronoi. Esses dois conceitos estão intimamente ligados, de maneira que, dada a triangulação induzida de Delaunay é possível obter o respectivo diagrama de Voronoi, e vice-versa.

O diagrama de Voronoi é uma estrutura geométrica que representa a informação de proximidade entre um conjunto de pontos. Dado um conjunto de pontos no plano, este é dividido em polígonos de Voronoi associados a cada ponto. O conjunto de todos esses polígonos forma o diagrama de Voronoi. Uma importante propriedade desse diagrama é que os pontos sobre ele têm a propriedade de serem equidistantes de dois ou mais pontos presentes no mesmo. O diagrama de Voronoi resolve o problema de determinar a área de abrangência de cada ponto de maneira otimizada. A triangulação de Delaunay consiste na interligação, através de segmentos de reta, dos pontos cujos polígonos de Voronoi são adjacentes. Essa estrutura tem a propriedade de que o círculo circunscrito em cada triângulo não contém nenhum ponto da triangulação. (SILVA e SOUSA, 2001).

Por outro lado, a triangulação induzida de Delaunay (Figura 2.16) é obtida sobrepondo-se a triangulação original a uma certa distribuição de probabilidade  $P(x)$ , de tal maneira que dois centros são conectados somente se os dois estão localizados em uma região onde  $P(x)$  é maior que zero (FRITZKE, 1995).

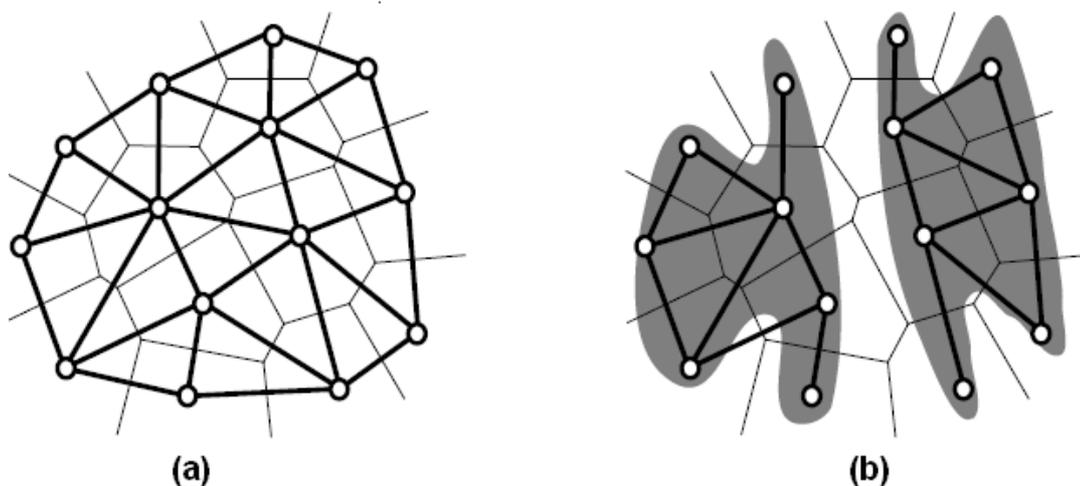


Figura 2.16 – (a) Diagrama de Voronoi (linhas mais finas) e a respectiva triangulação de Delaunay (linhas mais grossas); (b) Triangulação induzida de Delaunay, com as regiões sombreadas apresentando  $P(x) > 0$ .

Adaptado de (FRITZKE, 1995)

Dadas as amostras de entrada em  $R^k$ , cada nó  $n$  do grafo possui as seguintes características:

a) Um vetor *de pesos* de referência em  $R^k$ , chamado aqui de  $w_n$ .

Dado um nó qualquer do grafo, seu *vetor de pesos* (*weight vector*) pode ser interpretado como a posição desse nó no espaço vetorial das amostras.

b) Uma variável que contém o erro acumulado do nó, chamada aqui de  $error_n$ . A cada iteração do algoritmo durante o processo de criação do grafo, é medido a distância (segundo o tipo de distância utilizada) entre a amostra  $a_m$  sendo treinada e o nó  $n$  mais próximo. O erro existente (ou seja, a distância existente) entre a amostra  $a_m$  e o vetor  $w_n$  do nó mais próximo é acumulado na variável ( $error_n$ ) de erro desse nó. É através desses erros acumulados que o algoritmo vai decidir em qual posição

serão incluídos os novos nós, a medida em que o grafo está sendo criado. Depois de um certo número de iterações, o nó do grafo que possui o maior erro acumulado está sendo o representante de muitas amostras, e portanto o novo nó a ser adicionado deveria ser incluído próximo desse, dividindo o número de amostras representadas e com isso melhorando a distribuição das amostras entre todos os nós do grafo.

c) Uma lista de arestas contendo todas as ligações do nó  $n$  aos seus vizinhos. Cada uma das dessas arestas possui uma propriedade que é chamada de tempo de existência da aresta ou “idade”. Essa idade é utilizada para que o algoritmo possa remover as arestas mais antigas, e essa é uma das características que permitem ao algoritmo criar grafos cuja forma sendo modificada ao longo do treinamento, obedecendo à distribuição espacial das amostras de entrada.

O algoritmo possui os parâmetros listados a seguir. A medida que seus passos forem descritos, cada parâmetro será melhor compreendido.

- $\lambda$  – Número de iterações até que seja inserido um novo neurônio
- $e_b$  - Taxa de aprendizado para o neurônio vencedor
- $e_n$  - Taxa de aprendizado para os vizinhos do neurônio vencedor
- $\alpha$  - Taxa de desconto dos erros aplicada somente aos dois vencedores.
- $\beta$  - Taxa de desconto dos erros aplicada a todos os nós
- $a_{\max}$  - Peso máximo de uma aresta antes que ela seja removida

Cada um dos passos do algoritmo é descrito detalhadamente a seguir.

**Passo 0:** o primeiro passo consiste em obter aleatoriamente, segundo a distribuição de probabilidades do conjunto de amostras de entrada, uma amostra de entrada qualquer  $a_m$  (sinal de entrada). Além disso, são criados dois nós em posição qualquer no espaço  $R^n$ , o espaço das amostras de entrada. Esses dois nós devem ser ligados por uma aresta, a qual recebe 0 como idade, e o erro local dos dois nós criados é iniciado como 0.

**Passo 1:** encontrar dentre todos os nós já existentes do grafo os dois nós  $s$  e  $t$  cujos vetores de pesos  $w_s$  e  $w_t$  estiverem às duas menores distâncias da amostra  $a_m$ .

**Passo 2:** o nó  $s$ , o mais próximo da amostra  $a_m$ , é o nó vencedor e por isso terá valores atualizados. Para isso acumula-se a variável que mantém o erro do nó  $s$  ( $error_s$ ) de acordo com

$$error_s = error_s + distance(a_m, s)$$

onde  $distance(a_m, s)$  é uma função que retorna a distância de  $a_m$  a  $s$ .

**Passo 3:** mover  $s$  e seus vizinhos topológicos em direção a  $a_m$  de acordo com frações  $e_w$  e  $e_n$  da distância:

$$w_s = w_s + e_w * (a_m - w_s)$$

$$w_n = w_n + e_n * (a_m - w_n), \forall n \in Vizinhana(s)$$

**Passo 4:** incrementar de uma unidade a idade de todas as arestas que ligam o nó  $s$  aos seus vizinhos topológicos.

**Passo 5:** caso os nós  $s$  e  $t$  sejam ligados por uma aresta, essa aresta deve ter sua idade reiniciada. Além disso, caso essa aresta ainda não exista, ela deve ser criada.

**Passo 6:** caso exista alguma aresta com idade maior que a idade parâmetro  $a_{\max}$ , essa aresta deve ser removida do grafo. Depois dessa remoção todos os nós que estiverem desconectados, isto é, sem pelo menos um vizinho topológico, deverão ser removidos.

**Passo 7:** caso essa iteração do algoritmo seja um múltiplo de  $\lambda$ , e o critério de parada ainda não tiver sido atingido, um novo nó deverá ser inserido no grafo. Esse novo nó  $r$  será inserido da seguinte maneira:

- encontrar o nó  $u$  do grafo que possui o maior erro acumulado ( $error_u$ );
- encontrar, dentre todos os vizinhos de  $u$  o nó  $v$  com maior erro acumulado ( $error_v$ );
- criar o novo nó  $r$  e inseri-lo na distância média entre os nós  $u$  e  $v$ :

$$w_r = (w_u + w_v) / 2$$

- criar arestas ligando o nó  $r$  ao nó  $u$  e ao nó  $v$ , e depois remover a ligação existente entre  $u$  e  $v$ ;
- decrescer as variáveis de erro acumulado de  $u$  e  $v$  segundo o parâmetro  $\alpha$ , e inserir o valor do erro acumulado inicial do nó  $r$  como sendo o mesmo do nó  $u$ :

$$error_u = \alpha * error_u$$

$$error_v = \alpha * error_v$$

$$error_r = error_u$$

**Passo 8:** decrescer os valores das variáveis de erro de todos os nós do grafo de acordo com o parâmetro  $\beta$ ;

**Passo 9:** caso o critério de parada não seja atingido, voltar ao passo 1.

Uma explicação dos principais passos do algoritmo será feita a

seguir, para fins de compreensão do funcionamento do mesmo.

No passo 2, a variável de erro local acumulado do nó vencedor é atualizada. Essa variável mantém o valor do somatório das distâncias entre o nó e as amostras para as quais ele era o nó vencedor (situados à menor distância). Manter esse somatório é a forma encontrada para encontrar nós que estejam cobrindo uma porção relativamente grande da distribuição das amostras. Dessa forma, o erro local de cada nó é considerado uma importante medida estatística. Isso porque os nós que estejam cobrindo uma porção relativamente grande da distribuição das amostras de entrada terão, estatisticamente, seu erro total acumulado maior do que o outros nós. Quando nós cobrem grandes porções, ocorrem grandes somatórios de erros locais. Assim, como o objetivo do algoritmo é minimizar o erro total do grafo, saber quais são os nós que tem maior erro acumulado é um bom indicativo de em que áreas os novos nós devem ser incluídos durante o processo de crescimento do grafo.

No passo 3, uma vez encontrado o nó vencedor para cada amostra, existe uma movimentação desse nó, bem como de seus vizinhos topológicos, na direção da amostra segundo frações definidas por dois dos parâmetros fornecidos ao algoritmo,  $e_w$ , para o nó vencedor, e  $e_n$  para seus vizinhos topológicos. Isso é feito com o objetivo que, ao longo do treinamento, o nó tenha a todo momento seu vetor de pesos situado o mais próximo possível do centro do espaço compreendido pelo conjunto de amostras para as quais esse nó é o vencedor.

De acordo com Fritzke (1995), a inserção de ligações entre o vencedor e o segundo neurônio mais próximo da amostra gera uma simples conexão de uma triangulação induzida de Delaunay em relação à

posição atual de todos os nós. Assim, à medida em que novas ligações são criadas e removidas, o modelo tenta esboçar uma triangulação induzida de Delaunay, a qual é lentamente construída devido aos passos de adaptação dos vetores dos nós.

O decréscimo do somatório dos erros de todos os nós, ocorrido no passo 8 é explicado pelo fato de ser desejável que erros ocorridos recentemente tenham maior influência no somatório.

Os deslocamentos são lineares em relação à distância entre vetor de pesos do nó e a amostra, sendo que o valor do parâmetro que define o deslocamento do vencedor ( $e_w$ ) é geralmente muito maior que o valor que define os deslocamentos dos seus vizinhos ( $e_n$ ). Esses dois parâmetros variam de acordo com o problema em questão. Entretanto, valores relativamente altos tendem a favorecer a criação de grafos cuja posição dos nós é muito instável, variando fortemente no decorrer do treinamento, enquanto valores excessivamente baixos irão tornar a adaptação do grafo lenta e ineficaz.

No algoritmo GNG os nós são inseridos sempre após a execução de um determinado número  $\lambda$  de treinamentos. Ou seja, a cada  $\lambda$  treinamentos um novo nó é inserido no grafo, sendo sua posição de inserção definida pelo nó que possui o maior erro acumulado no momento. É dessa forma que se consegue manter uma boa distribuição dos nós do grafo em relação a distribuição espacial das amostras de treino.

O parâmetro  $\lambda$  tem grande impacto na performance do algoritmo. Caso esse parâmetro tenha seu valor excessivamente baixo a distribuição inicial dos nós no grafo será mal executada, pois ao se incluir um novo nó os nós existentes não terão seu vetor de pesos situados em posições

realmente representativas da distribuição amostral. No entanto, escolhas de valor alto para  $\lambda$  farão com que o tempo de execução do algoritmo seja aumentado, uma vez que serão necessários vários treinamentos antes da inclusão de cada novo neurônio.

Como critérios de parada do algoritmo, pode-se escolher um número pré-definido de *clusters* a serem gerados, ou mesmo escolher que o processo de agrupamento simplesmente seja interrompido assim que os critérios de erro sejam atingidos.

A Figura 2.17 exemplifica a evolução do algoritmo GNG. Inicialmente, o algoritmo conta com dois neurônios inicializados em posições aleatórias (Figura 2.17-a). À medida que aumentam o número de iterações, os neurônios vão se ajustando às regiões onde  $P(x) > 0$ . As linhas amarelas correspondem a ligações da triangulação induzida de Delaunay, enquanto que linhas vermelhas dão origem às regiões de Voronoi representadas por cada neurônio.

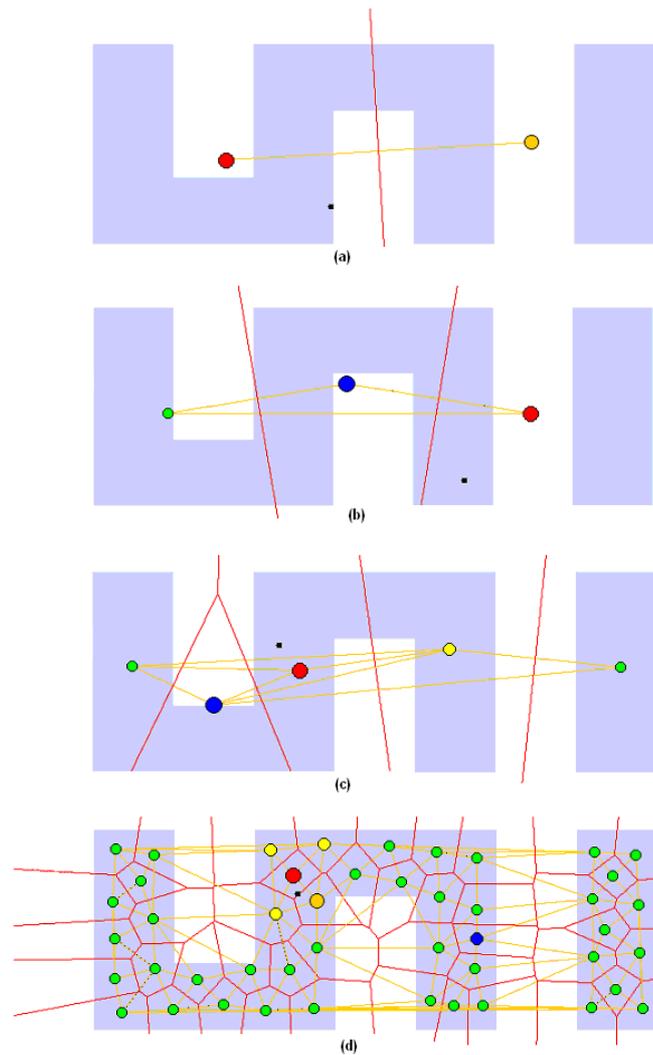


Figura 2.17 – Evolução do algoritmo GNG. A região sombreada corresponde a áreas onde  $P(x) > 0$ . As linhas amarelas correspondem a ligações da triangulação induzida de Delaunay; linhas vermelhas formam as regiões de Voronoi de cada neurônio. Fonte: Institut für Neuroinformatik (2006).

## 2.6 Máquinas de Vetor de Suporte

A Máquina de Vetor de Suporte (MVS) introduzida por Vapnick (1995) é um método de aprendizagem supervisionada usado para estimar uma função que classifique dados de entrada em duas classes. A idéia

básica por trás da MVS é construir um hiperplano como superfície de decisão, de tal maneira que a margem de separação entre as classes seja máxima. O objetivo do treinamento através de MVS é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização.

As MVS são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Os algoritmos de treinamento das MVS possuem forte influência da teoria de otimização e de aprendizagem estatística. Em poucos anos, as MVS vêm demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações (CRISTIANINI e SHAW, 2000).

Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador MVS é capaz de encontrar um hiperplano baseado em um conjunto de pontos denominados “vetores de suporte”, o qual maximiza a margem de separação entre as classes. Por hiperplano entende-se uma superfície de separação de duas regiões num espaço multidimensional, onde o número de dimensões possíveis pode ser, até, infinito. Mesmo quando as duas classes não são separáveis, a MVS é capaz de encontrar um hiperplano através do uso de conceitos pertencentes à teoria da otimização. A Figura 2.18, mostra em duas dimensões para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha mais escura), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

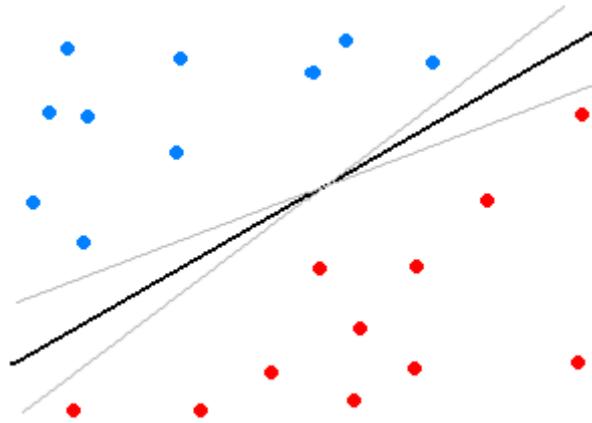


Figura 2.18 - Separação de duas classes através de hiperplanos

Seja o conjunto de amostras de treinamento  $(x_i, y_i)$ , sendo  $x_i \in \mathfrak{R}^n$  o vetor de entrada,  $y_i$  a classificação correta das amostras e  $i = 1, \dots, n$  o índice de cada ponto amostral. O objetivo da classificação é estimar a função  $f : \mathfrak{R}^n \rightarrow \{\pm 1\}$ , que separe corretamente os exemplos de teste em classes distintas.

A etapa de treinamento, tenta estimar a função  $f(x) = (w \cdot x) + b$ , procurando por valores de  $w$  e  $b$  tais que a seguinte relação seja satisfeita

$$y_i((w \cdot x_i) + b) \geq 1 \quad (2.11)$$

Os valores ótimos de  $w$  e  $b$  serão encontrados ao minimizar a seguinte equação, de acordo com a restrição dada por 2.11 (CHAVES, 2006).

$$\Phi(w) = \frac{w^2}{2} \quad (2.12)$$

MVS ainda possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação nos casos em que uma perfeita separação entre as duas classes não for possível. Isso graças a inclusão de variáveis de folga, que permitem que as restrições presentes em 2.11

sejam quebradas.

O problema de otimização passa a ser então a minimização da Equação 2.13, de acordo com a restrição imposta pela Equação 2.11.  $C$  é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento e deve ser selecionado pelo usuário.

$$\Phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^N \xi_i \quad (2.13)$$

$$y_i((w \cdot x_i) + b) + \xi_i \geq 1 \quad (2.14)$$

Através da teoria dos multiplicadores de Lagrange, chega-se à Equação 2.15. O objetivo então passa a ser encontrar os multiplicadores de Lagrange  $\alpha_i$  ótimos que satisfaçam a Equação 2.16 (CHAVES, 2006).

$$w(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j (x_i x_j) \quad (2.15)$$

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \quad (2.16)$$

Apenas os pontos onde a restrição 2.10 seja exatamente igual à unidade têm correspondentes  $\alpha_i \neq 0$ . Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos têm fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento.

A Figura 2.19 destaca os pontos que representam os vetores de suporte. Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

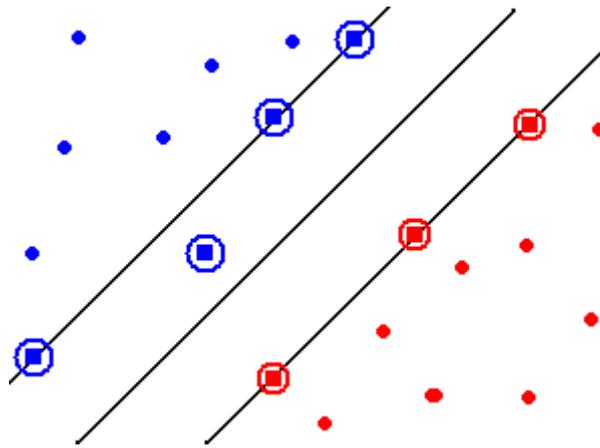


Figura 2.19 - Vetores de suporte (destacados por círculos)

Para que a MVS possa classificar amostras que não são linearmente separáveis, é necessário uma transformação não-linear que transforme o espaço entrada (dados) para um novo espaço (espaço de características). Esse espaço deve apresentar dimensão suficientemente grande, e através dele, a amostra pode ser linearmente separável. Dessa maneira, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de características ao invés do espaço de entrada original. Essa construção depende do cálculo de uma função  $K$  de núcleo de um produto interno (HAYKIN, 2001). A função  $K$  pode realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

A Equação 2.17 mostra o resultado da Equação 2.15 com a utilização de um núcleo.

$$w(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2.18)$$

Uma importante família de funções de núcleo é a função de base radial, muito utilizada em problemas de reconhecimento de padrões e

também utilizada neste trabalho. A função de base radial é definida por

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \quad (2.19)$$

Neste trabalho, a MVS participou das etapas de classificação dos candidatos a massa.

## 2.7 Medidas de Desempenho

Esta seção examina as ferramentas utilizadas para testar a eficácia dos métodos propostos neste trabalho.

Em problemas de reconhecimento de padrões ligados à área médica, costuma-se medir o desempenho da metodologia calculando-se algumas estatísticas descritivas, tais como sensibilidade e especificidade. Sensibilidade é a proporção de verdadeiro-positivos corretamente identificados pelo teste, definida por  $VP/(VP + FN)$ , onde  $VP$  é o número de verdadeiro-positivos e  $FN$  o número de falso-negativos. Especificidade é a quantidade de verdadeiro-negativos corretamente identificados pelo teste, e é definida por  $VN/(VN + FP)$ , onde  $VN$  é o número de verdadeiro-negativos e  $FP$  o número de falso-positivos. A taxa de classificação correta ou acurácia ( $A$ ) é definida como a razão entre o número de casos na amostra em estudo que foram classificados corretamente e o número total de casos na amostra em estudo (AZPITARTE, 2006).

Embora possam existir várias medidas para verificar o desempenho de um classificador, a medida mais importante é o desempenho do mesmo a partir da classificação de novos casos (conjunto

de testes). O desempenho do classificador, medido através do conjunto de teste, é uma boa indicação de seu desempenho real.

Um método simples para avaliar o rendimento de um sistema de detecção de objetos em imagens consiste em determinar o número de falsos positivos que se produzem a cada imagem. Assim, a partir do número total de falsos positivos obtidos durante os testes, obtém-se a taxa de falsos positivos por imagem (FP/Imagem) dividindo-se esse total pelo número de imagens utilizadas.

### 3 METODOLOGIA

Este capítulo descreve os procedimentos utilizados para a detecção de massas em mamografias digitalizadas. Primeiramente será examinada a base de imagens utilizadas nos testes. A seguir, será descrita a seqüência de ações adotadas para a detecção de massas nas imagens, através da segmentação de objetos de interesse nas imagens, a redução dos candidatos a massa, além da classificação de cada objeto em massa ou não-massa através de Máquinas de Vetor de Suporte (MVS).

#### 3.1 Amostra

Este trabalho utilizou amostra de mamografias digitalizadas obtidas através da base de dados DDSM (*Digital Database for Screening Mamography*), disponível gratuitamente na internet (HEATH *et al.*, 1998). Essa base contém 2620 casos adquiridos através das seguintes instituições americanas: *Massachusetts General Hospital*, *Wake Forest University*, e *Washington University in St. Louis School of Medicine*.

Cada caso contém duas imagens de cada mama, nas projeções médio-lateral oblíqua e crânio-caudal. Além disso, são disponibilizadas informações sobre a paciente, tal como a idade e a densidade da mama. Também são informados o tipo de *scanner* utilizado na digitalização do exame e a resolução da imagem. Imagens com áreas suspeitas possuem a descrição da anormalidade, o diagnóstico e a localização da mesma na imagem. É importante frisar que tanto a localização quanto o diagnóstico

das massas utilizadas neste trabalho foram realizadas por um especialista (HEATH *et al.*, 1998).

Neste trabalho foram utilizadas 997 imagens de mamografias pertencentes à base DDSM, contendo pelo menos uma massa. Desse total, 436 foram separadas para o treinamento e teste dos classificadores MVS e 561 foram utilizadas para testar a eficácia do processo de detecção de massas em mamografias digitalizadas.

### **3.2 Software e hardware utilizados**

Para a implementação dos métodos descritos neste trabalho utilizou-se a linguagem C++, através da ambiente de desenvolvimento *freeware* Dev-C++.

A manipulação das imagens foi realizada através das funções da biblioteca CImg (TSCHUMPERLÉ, 2007), disponível gratuitamente na internet. A biblioteca pode ser usada com a linguagem C++, fornecendo classes simples e funções para carregar, salvar, processar e exibir imagens. Possui grande portabilidade, funcionando perfeitamente em ambientes Unix/X11, Windows, MacOS X e sistemas operacionais FreeBSD.

Através da biblioteca LIBSVM (CHANG e LIN, 2007) foram implementados os classificadores MVS utilizados neste trabalho. A biblioteca encontra-se disponível gratuitamente na internet.

O computador utilizado para os testes foi um AMD Sempron 1.8 GHz, com 512 MB de memória RAM e HD com 60 GB.

### 3.3 Metodologia proposta

A Figura 3.1 apresenta a seqüência de etapas utilizada neste trabalho para a detecção de massas em mamografias digitalizadas.

Inicialmente, a imagem é pré-processada. Esta etapa consiste em remover da imagem estruturas que não pertençam à mama, com o intuito de processar apenas a parte que realmente interessa na detecção de massas em mamografias.

Em seguida, a imagem resultante tem suas estruturas semelhantes separadas através do algoritmo *Growing Neural Gas* (Seção 2.5). A saída dessa etapa consiste em um conjunto de objetos segmentados da mama, os quais são considerados candidatos a massa.

Para reduzir os candidatos a massa obtidos na etapa anterior, cada objeto separado tem suas medidas geométricas (Seção 2.4) extraídas. Essas medidas (excentricidade, circularidade, compacidade, desproporção circular e densidade circular) são submetidas à avaliação de um classificador MVS previamente treinado e então classificadas em massa ou não-massa de acordo com a geometria. Essa etapa tem por objetivo de descartar antecipadamente objetos alongados, “ocos” ou “retorcidos” demais. Assim, somente os objetos classificados como pertencentes à classe “massa” passa para a etapa seguinte.

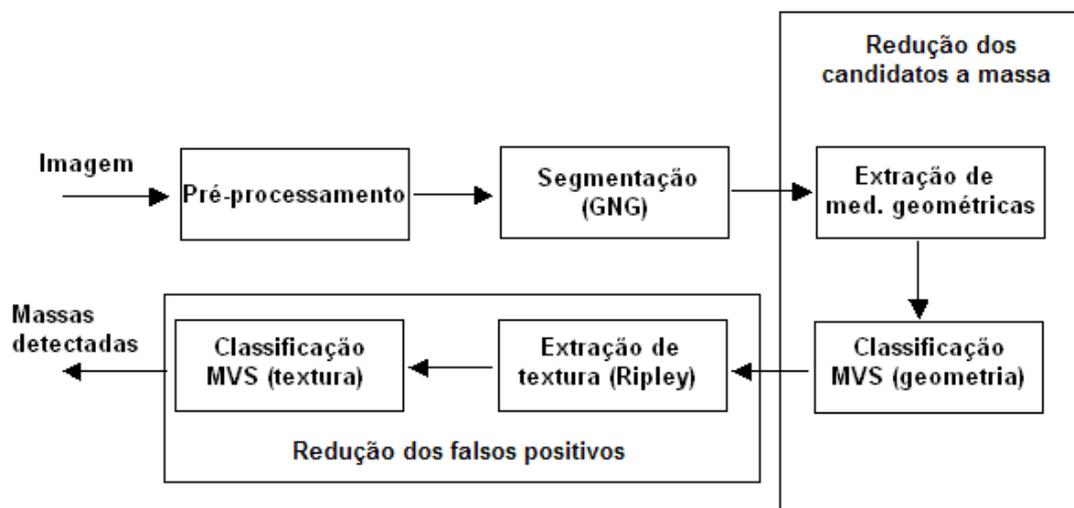


Figura 3.1 – Metodologia proposta

O próximo estágio corresponde à extração de textura dos objetos através da modificação da função  $K$  de Ripley examinada na Seção 2.3.2 e a classificação de cada objeto em massa ou não-massa através de um classificador MVS.

A seguir, será feita uma descrição detalhada de cada etapa acima citada, bem como os parâmetros utilizados em cada método.

### 3.3.1 Pré-processamento

Antes da execução do algoritmo de agrupamento, é preciso garantir que a imagem que será tratada pelo algoritmo possua apenas a estrutura que interessa, ou seja, apenas a mama em si. Em uma mamografia (Figura 2.1) geralmente são adicionados rótulos que trazem informações como o tipo de projeção (médio-lateral oblíqua ou crânio-caudal), a data do exame ou ainda a lateral do corpo a qual pertence a mama (esquerda ou direita).

Dessa forma, neste trabalho foi utilizado um método bastante simples para segmentar apenas a parte da imagem correspondente à mama. Inicialmente, é criada uma cópia da imagem original. Essa cópia é então binarizada com base em um limiar, ou seja, se um *pixel* possui valor menor que o limiar, esse *pixel* recebe valor 0; caso contrário, o *pixel* recebe valor 255 (valor máximo permitido em uma imagem em tons de cinza, correspondente à cor branca), de acordo com a Seção 2.2.3. Diversos valores de limiares foram testados, entretanto, melhores resultados foram obtidos utilizando-se o valor de 80 para binarizar a imagem.

Se a imagem representar uma mama direita, por exemplo, o algoritmo inicializa no canto superior esquerdo da imagem. Em caso contrário, a inicialização da varredura ocorre no canto superior direito. Dessa maneira, de acordo com a especificação do tipo de mama (esquerda ou direita), o algoritmo varre a imagem, para cada linha, até encontrar um *pixel* diferente de branco. Quando um *pixel* de valor 0 é encontrado, o restante da linha da imagem original é zerado, apagando quaisquer estruturas não pertencentes à mama. A Figura 3.2 mostra duas imagens de mamografia. A primeira corresponde à imagem original, com objetos que não pertencem à mama em si. A segunda é a imagem após a execução do algoritmo para remoção desses objetos.

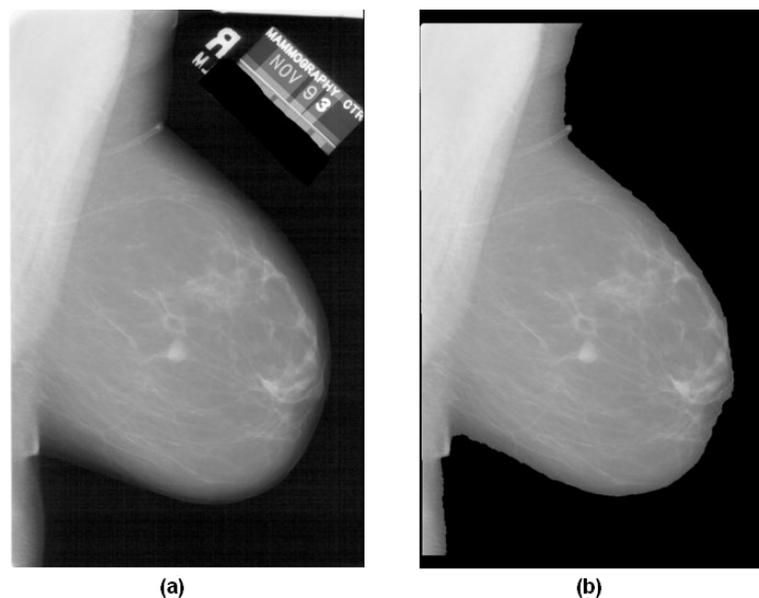


Figura 3.2 – Imagem de mamografia antes da segmentação da mama (a);  
imagem após a segmentação (b)

### 3.3.2 Segmentação

A etapa de segmentação consiste em agrupar as estruturas da imagem que apresentem propriedades semelhantes. Neste trabalho, o algoritmo *Growing Neural Gas* foi utilizado com esse propósito.

Em um processo de segmentação, em primeiro lugar, é necessário definir quais os elementos da imagem que serão comparados entre si e agrupados. Assim, a imagem foi dividida em janelas de tamanho  $4 \times 4$  pixels, com o intuito de extrair as características estatísticas mais relevantes de cada uma e para que, dessa maneira, cada janela correspondesse a um elemento a ser passado para o algoritmo de agrupamento. Diversos tamanhos de janelas foram testados, como por exemplo,  $2 \times 2$  e  $8 \times 8$ , entretanto, janelas  $4 \times 4$  demonstraram melhor

qualidade nos resultados. A Figura 3.3 mostra uma imagem de mamografia dividida em janelas 8x8.

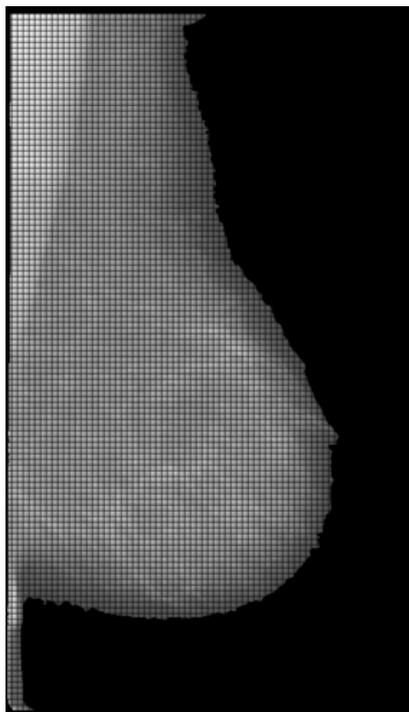


Figura 3.3 – Imagem dividida em janelas.

A informação estatística a ser extraída de cada objeto tem papel fundamental no desempenho do algoritmo de agrupamento. Neste trabalho, optou-se por extrair informações de maneira bastante simples. Os *pixels* de cada janela foram armazenados em um vetor de 16 posições e ordenados, do menor para o maior. Dessa maneira, cada janela é representada em um vetor de características no qual o primeiro elemento corresponde ao menor valor e o último elemento ao maior valor, indicando ao algoritmo a variação de *pixels* presentes em cada janela.

O uso deste vetor de características em particular apresenta algumas vantagens. A informação presente em cada janela é na maioria das vezes homogênea e bastante limitada. Testes empíricos demonstraram que o uso de medidas de textura para a descrição estatística de cada

janela mostraram resultados abaixo do esperado na tarefa de segmentação das estruturas da mama, além de demandar grande período de tempo na execução dessa tarefa. Assim, optou-se por um processo mais simples, mas que mostrasse a variação de pixels presentes em cada janela. Os testes empíricos realizados com esse vetor de características específico demonstraram os melhores resultados.

A amostra ou sinal de entrada em cada iteração do algoritmo é obtido através da escolha aleatória de um vetor de características pertencentes a uma janela da imagem. Dessa maneira, o algoritmo consegue refletir a topologia dos dados de entrada, através do processo de adaptação descrito na Seção 2.5.1.

Através de testes empíricos, foram encontrados os seguintes valores para os parâmetros do algoritmo GNG.

- $\lambda = 1000$
- $e_b = 0,5$
- $e_n = 0,005$
- $\alpha = 0,5$
- $\beta = 0,95$
- $a_{\max} = 50$

Assim que o número de neurônios existentes se torna igual ao número de grupos desejados, o algoritmo pára a sua execução. Dessa maneira, esta implementação do algoritmo apresenta um parâmetro extra, indicado pelo número máximo de neurônios permitido, que é igual ao número de grupos desejado. Essa abordagem é a mesma adotada no trabalho de Figueiredo (2006).

Neste trabalho, adotou-se a estratégia de variar o número de

grupos desejados, como forma de separar estruturas da imagem que, com determinado número de grupos, não poderiam ser separadas. Assim, o algoritmo foi executado para separar a imagem em 7, 8 ou 9 grupos. Isso porque foi verificado que muitas vezes o algoritmo não é capaz de separar a massa correta utilizando 7 grupos, porém é capaz de separar a mesma utilizando 8 ou 9 grupos, ou vice-versa. É importante ressaltar que, no processo de detecção, é feita a união de todos os objetos segmentados usando 7, 8 e 9 grupos, sendo examinados todos esses objetos. A Figura 3.4 mostra uma mamografia que foi segmentada utilizando esses valores de grupos.

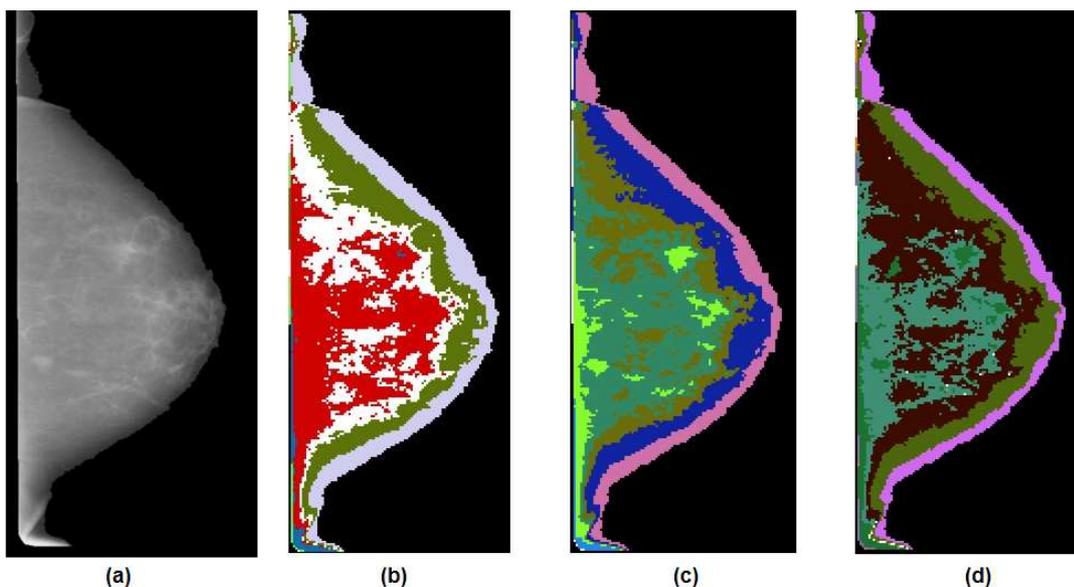


Figura 3.4 – Mamografia segmentada através do algoritmo GNG. (a) Imagem original; (b) utilizando 7 grupos; (c) utilizando 8 grupos; (d) utilizando 9 grupos

### 3.3.3 Redução dos candidatos a massa

Esta etapa tem por objetivo a eliminação de objetos alongados, “ocos” ou “retorcidos” demais, os quais não correspondem às formas de massa examinadas na Seção 2.1.1.

A base de imagens usada para treinamento e teste do classificador MVS é composta de objetos segmentados através do algoritmo *Growing Neural Gas*. Inicialmente, são descartados objetos com largura e altura maiores que 300 *pixels* bem como aqueles considerados muito pequenos, na faixa de 16x16 *pixels*. Isso porque foi verificado que para a base de imagens utilizada neste trabalho, as dimensões das massas encontram-se nesse intervalo.

A partir da base de objetos segmentados, um subconjunto dessa base é selecionado para treinamento de um classificador MVS, a partir de suas medidas geométricas. As medidas geométricas extraídas são obtidas a partir dos objetos binarizados. Tais medidas são excentricidade, circularidade, compacidade, desproporção circular e densidade circular, descritas na Seção 2.4.

Assim, de acordo com a metodologia proposta na Figura 3.1, esse classificador será responsável por filtrar as amostras, fazendo com que apenas os objetos que são indicados pelo mesmo como “massa” passem à próxima fase. Ou seja, somente os objetos indicados como “massa”, de acordo com a geometria, fazem parte das amostras que serão examinadas na classificação baseada em textura. Dessa maneira, esta etapa funciona como uma redução dos candidatos a massa, fornecendo à próxima etapa apenas objetos cuja descrição geométrica se assemelha o

máximo possível à forma esperada de uma massa. A Figura 3.5 exemplifica alguns objetos dessa base de amostras.

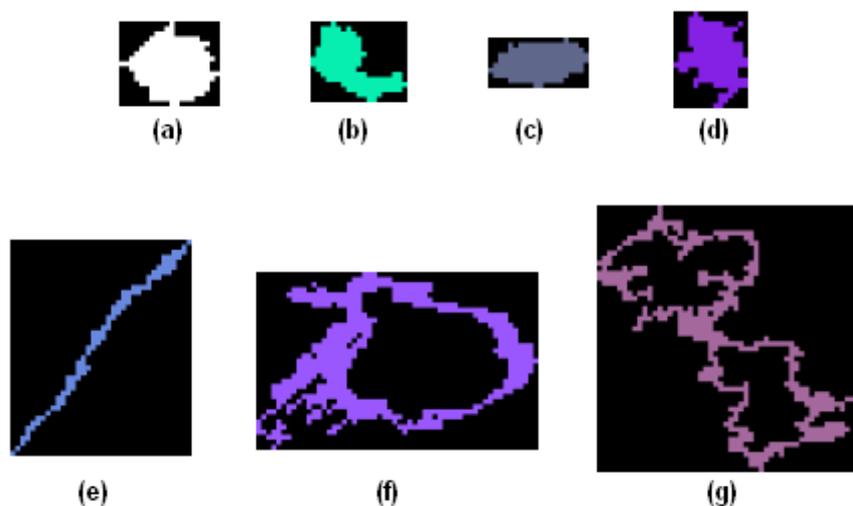


Figura 3.5 – Objetos considerados “massa” (a,b,c,d) e “não-massa” (e,f e g), de acordo com sua geometria

Os parâmetros do classificador MVS utilizado para o treinamento foram  $C = 2$  e  $\gamma = 3 \times 10^{-5}$ . A função de núcleo utilizada foi a função de base radial, definida na Seção 2.6.

### 3.3.4 Redução dos falsos positivos

Os objetos que são classificados como massa na etapa anterior têm as suas características de textura extraídas e são submetidos à avaliação de um classificador MVS com o intuito de determinar se o mesmo é ou não uma massa.

Neste trabalho, utilizou-se a modificação proposta na Seção 2.3.2 para a função  $K$  de Ripley, com o objetivo de extrair a textura dos

objetos de interesse.

O centro da função K de Ripley local tem papel importante na descrição de textura de um objeto segmentado, pois a partir dele são calculadas as variações de textura ao longo da distância. Para cada objeto, o centro escolhido para o cálculo da função K local foi o centro de gravidade (ou centro de massa) da figura. A Figura 3.6 mostra a localização do centro de massa para diferentes objetos. As coordenadas  $x_c$  e  $y_c$  desse ponto podem ser determinadas através das equações (MARSHALL, 1997)

$$x_c = \frac{1}{n} \sum_x x \quad (3.1)$$

$$y_c = \frac{1}{n} \sum_y y \quad (3.2)$$

Sendo  $n$  o número de pontos pertencentes ao objeto.



Figura 3.6 – Localização do centro de massa

Foram utilizados 3 raios para a análise da variação de textura ao longo do objeto. Isso porque o número de raios utilizado está diretamente relacionado com o número de variáveis repassadas ao classificador MVS. Assim, com o intuito de simplificar o modelo do classificador, procurou-se utilizar poucos raios, mas de tal maneira que a separação das classes não fosse comprometida. Através de sucessivos testes, o valor 3 foi considerado aceitável para esse fim.

Para a determinação dos três raios em cada objeto, é necessário

calcular o raio máximo  $R_{max}$  que pode ser alcançado a partir do seu centro, ou seja a distância entre o centro de massa da figura até o *pixel* mais distante pertencente à mesma. Cada raio é obtido através da seguinte equação

$$R_i = \frac{R_{Max}}{3} * i \quad \text{para } i = 1,2,3 \quad (3.3)$$

Dessa maneira, a área de cada anel concêntrico utilizada no cálculo da função  $K$  de Ripley modificada é obtida calculando-se a área do círculo de interesse (maior) menos a área do círculo anterior (menor), lembrando que  $R_0 = \frac{R_{Max}}{3} * 0 = 0$ .

A equalização do histograma, descrita na Seção 2.2.2, é realizada para cada elemento segmentado, com o objetivo de melhorar o contraste do objeto, e, conseqüentemente, prover uma melhor descrição da textura do mesmo (Figura 3.7).

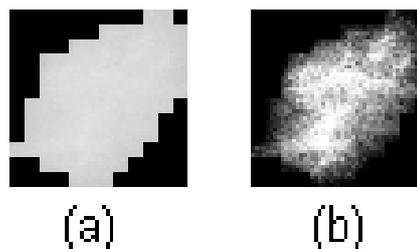


Figura 3.7 – Exemplo de uma massa segmentada. (a) - Antes do processo de equalização de histograma. (b) - Depois do processo de equalização de histograma

Cada objeto é quantizado para 8 níveis de cinza. Esse procedimento é necessário para reduzir a quantidade de variáveis do

modelo, ao mesmo tempo em que é preservada a descrição da textura à medida em que se varia a distância em relação ao centro. Dessa maneira, são usadas 24 variáveis (3 raios x 8 níveis de cinza) na classificação MVS.

A etapa final da metodologia proposta consiste em classificar cada objeto em massa ou não-massa de acordo com a textura obtida através da função  $K$  de Ripley modificada, descrita na Seção 2.3.3. Para isso foi utilizado um classificador MVS para separar os objetos em massa ou não-massa, de acordo com a descrição de sua textura.

A base de imagens usada para treinamento e teste do classificador MVS é composta de objetos segmentados através do próprio algoritmo *Growing Neural Gas*. Cada objeto teve seu histograma equalizado, assim como suas medidas de textura extraídas através da função  $K$  de Ripley local modificada (Seção 2.3.3). Além disso, cada um deles foi rotulado de acordo com a sua classe (massa ou não-massa).

Os parâmetros do classificador MVS utilizado nesta etapa foram  $c = 8$  e  $\gamma = 2$ . Novamente, a função de núcleo utilizada foi a função de base radial, definida na Seção 2.6.

Para a validação da classificação, novamente as amostras foram divididas em dois grupos, um para treinamento e o outro para testes.

Dessa maneira, ao final do processo, a metodologia proposta é capaz de identificar possíveis massas presentes em uma mamografia digital. A Figura 3.8 mostra um exemplo de detecção bem-sucedida, com a massa detectada indicada em vermelho.

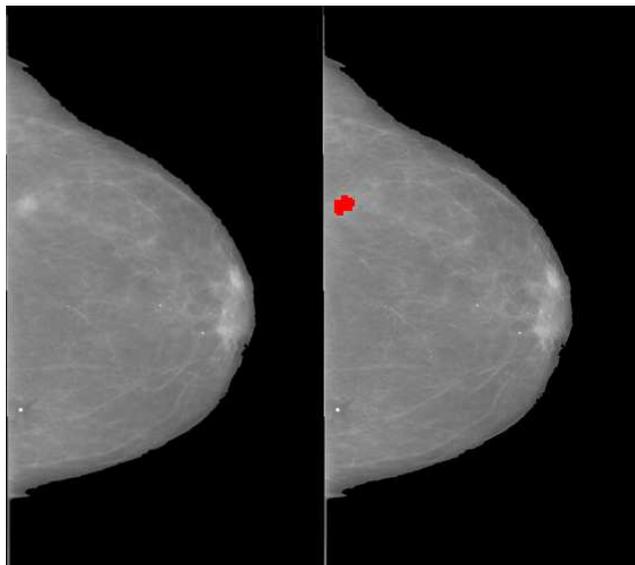


Figura 3.8 – Resultado final do processo de detecção de massas, indicando a localização correta da massa, em vermelho.

No próximo capítulo serão apresentados os resultados obtidos pela metodologia adotada neste trabalho.

## 4 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir do uso da metodologia proposta. São examinados o desempenho das principais etapas presentes na metodologia descrita no Capítulo 3. Para tanto, os resultados são examinados em duas etapas distintas. A primeira etapa é referente ao treinamento e teste dos classificadores MVS, para a redução dos candidatos a massa e redução dos falsos positivos. A segunda etapa diz respeito ao teste do processo de detecção propriamente dito. Ou seja, a partir do processamento de novas imagens de mamografias, de acordo com a metodologia proposta, é verificado o total de imagens cuja massa foi corretamente detectada, além do número de falsos positivos e falsos negativos por imagem.

Para a obtenção dos resultados mostrados neste capítulo foram utilizadas 997 imagens de mamografias contendo pelo menos uma massa. Desse total de imagens, 436 foram separadas para o treinamento e teste dos classificadores MVS e 561 foram usadas para avaliar o processo de detecção de massas em mamografias digitalizadas.

### 4.1 Treinamento e teste dos classificadores MVS

Para o treinamento e teste (validação) dos classificadores foram utilizadas 436 imagens de mamografias, contendo cada uma pelo menos uma massa. Essas imagens foram pré-processadas e segmentadas através do algoritmo *Growing Neural Gas* com 7, 8 e 9 grupos, o que resultou em vários objetos do tipo massa e não-massa. Foram

selecionados então 990 objetos, os quais constituem a base utilizada para treinamento e teste dos classificadores MVS. Desse total, 450 objetos são do tipo massa e 540 são do tipo não-massa.

A partir dessa base de objetos, utilizou-se um subconjunto da mesma para treinar o classificador MVS responsável pela etapa de redução dos candidatos a massa. Para isso, foram extraídas as medidas geométricas discutidas na Seção 2.4, as quais foram passadas ao classificador. Esse classificador treinado foi utilizado para reduzir os candidatos a massa na base inicial, contendo os 990 objetos, de maneira que somente os objetos classificados como massa, de acordo com a geometria, fizessem parte da base de objetos utilizada para a classificação de acordo com a textura.

#### **4.1.1 Redução dos candidatos a massa**

A partir da base inicial, foi selecionado um subconjunto de objetos que melhor descrevem cada classe. Aqui, parte-se do pressuposto que objetos da classe “massa” são em geral mais arredondados e circulares do que objetos do tipo não-massa. Estes, nessa etapa, são formados por objetos alongados, ocos e/ou retorcidos. Isso por que nesta etapa o objetivo é tão somente filtrar os objetos, descartando aqueles que apresentam características indesejáveis (alongados ou retorcidos demais) e deixando passar aqueles que são mais arredondados e circulares.

Dessa maneira, a base utilizada para treinamento e teste do classificador consiste de 295 objetos, os quais são um subconjunto da base de dados inicial, formada por 990 objetos. Desses 295 objetos, 120 são

massas e 175 são do tipo não-massas. O treinamento utiliza 96 amostras do tipo massa e 140 do tipo não-massa, os quais correspondem a 80% do valor total de cada classe. Para os testes, foram utilizadas 24 massas e 35 não-massas.

A Tabela 4.1 demonstra os resultados obtidos para o treinamento do classificador MVS.

Tabela 4.1 – Resultados para o treinamento (classificação baseada em geometria)

<b>Classe</b>	<b>Massa</b>	<b>Não-massa</b>
<b>Massa</b>	94	2
<b>Não-massa</b>	2	138
<b>Corretos (%)</b>	97,92	98,57

Os resultados dos testes são mostrados na Tabela 4.2.

Tabela 4.2 – Resultados para os testes (classificação baseada em geometria)

<b>Classe</b>	<b>Massa</b>	<b>Não-massa</b>
<b>Massa</b>	20	4
<b>Não-massa</b>	3	32
<b>Corretos (%)</b>	83,33	91,43

Observa-se que a classificação MVS, utilizando as medidas geométricas propostas, apresenta um bom desempenho, o qual é medido pela capacidade de generalização da aprendizagem demonstrada na predição das amostras de teste. Para os testes, a classificação obteve 83,33% de sensibilidade e 91,43% de especificidade, o que significa que as medidas geométricas propostas fornecem um bom suporte para a caracterização da geometria dos objetos segmentados. Além disso, percebe-se que a classificação realizada através de MVS apresenta um poder de generalização satisfatório para o problema em questão.

Aplicando o classificador treinado na base de dados inicial, obtém-se os objetos considerados como “massa”, de acordo com a

geometria. Assim, dos 990 objetos da base inicial, 743 foram selecionados como objetos de interesse para a próxima etapa. Isso caracteriza o processo de redução dos candidatos a massa, de acordo com a metodologia descrita na Figura 3.1.

É importante ressaltar que essa classificação tem por objetivo apenas reduzir os candidatos a massa. Ou seja, o objetivo aqui é que o classificador atue como um filtro que corte os objetos cuja forma geométrica é excêntrica demais, o que não corresponde ao que se espera de uma massa. Assim somente os objetos de interesse passarão à próxima fase, na qual são definidos de fato quais objetos são massa com base na textura, calculada a partir da função  $K$  de Ripley.

#### **4.1.2 Redução dos falsos positivos**

Os objetos marcados como massa na classificação anterior têm suas características de textura extraídas para fazer parte do treinamento e teste do segundo classificador MVS. A base utilizada para treinamento e teste de 743 objetos, os quais foram obtidos a partir da classificação anterior (baseada em medidas geométricas). Tal base é composta de 367 massas e 376 não-massas. O treinamento utiliza 288 amostras do tipo massa e 306 do tipo não-massa, o que corresponde a aproximadamente 80% dos valores totais de cada classe. Para os testes, foram utilizadas 79 massas e 70 não-massas.

A Tabela 4.3 demonstra os resultados obtidos para o treinamento do classificador MVS.

Tabela 4.3 – Resultados para o treinamento (classificação baseada em textura)

<b>Classe</b>	<b>Massa</b>	<b>Não-massa</b>
<b>Massa</b>	277	34
<b>Não-massa</b>	11	272
<b>Corretos (%)</b>	96,18	88,89

A Tabela 4.4 apresenta o resultado dos testes realizados através do classificador MVS para validar a classificação.

Tabela 4.4 – Resultados para os testes (classificação baseada em textura)

<b>Classe</b>	<b>Massa</b>	<b>Não-massa</b>
<b>Massa</b>	67	10
<b>Não-massa</b>	12	60
<b>Corretos (%)</b>	84,81	85,71

A análise dos resultados dessa etapa permite inferir que a classificação MVS em conjunto com a função K de Ripley modificada apresenta um bom desempenho, o qual também pode ser medido pela capacidade de generalização demonstrada durante os testes (Tabela 4.4). Para os testes, a classificação obteve 84,81% de sensibilidade e 85,71% de especificidade. A partir disso, é possível inferir que as medidas de textura propostas neste trabalho constituem uma boa ferramenta para a descrição da textura dos objetos segmentados. Novamente, os resultados indicam que a classificação realizada através de MVS apresenta um poder de generalização satisfatório para o problema em questão.

## 4.2 Resultados da detecção

Nesta seção é examinada a eficácia da metodologia proposta como um todo. Foram utilizadas para esse propósito 561 novas imagens de mamografia, contendo pelo menos uma massa. Assim, para cada imagem

de mamografia, são seguidos todos os passos da metodologia, descritos na Figura 3.1. Para isso, utiliza-se o algoritmo de limiarização para a separação da mama, o algoritmo GNG para obtenção dos candidatos a massa, os classificadores MVS previamente treinados para a redução dos candidatos a massa e para a redução dos falsos positivos.

Para medir a eficácia da segmentação de anormalidades em imagens médicas, é necessária a avaliação de um especialista ou técnico em radiografias, capaz de indicar se o algoritmo separou ou não a região suspeita do restante da imagem. A base de imagens DDSM, utilizada neste trabalho, disponibiliza a localização da anormalidade nos casos em que a mesma existe (Figura 4.1).

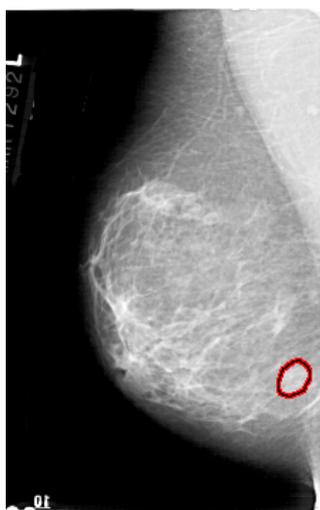


Figura 4.1 – Localização de uma massa. Fonte: *Digital Database for Screening Mammography – DDSM*

O algoritmo obteve sucesso em segmentar a massa em 509 das 561 imagens utilizadas para testar o processo de detecção, o que significa um percentual de 90,73% de acerto na segmentação da massa. A Figura 4.2 mostra uma mamografia na qual o algoritmo obteve êxito na

segmentação da massa do restante da imagem.

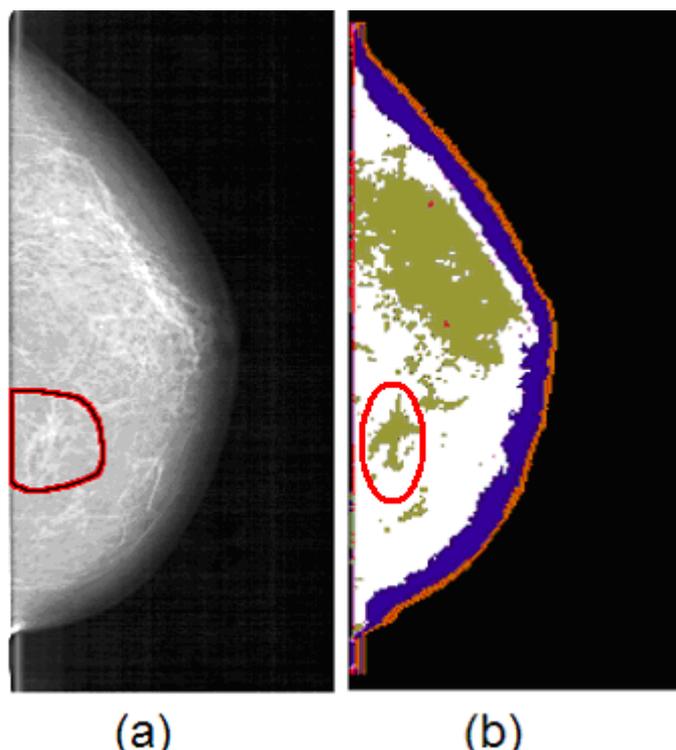


Figura 4.2 – Segmentação correta de uma massa. (a) - Imagem fornecida pelo DDSM ; (b) Imagem segmentada pelo GNG. A massa nos dois casos está destacada em vermelho.

Através do uso da metodologia proposta na Seção 3.3, e utilizando o modelo de classificação MVS cujos resultados são descritos na Seção 4.1, a metodologia detectou massas em 501 imagens, de um total de 561, o que caracteriza uma taxa de acerto de 89,30%. Com relação aos falsos positivos, foram indicados como sendo massas 519 objetos que correspondiam a tecidos normais. Com isso, a presente metodologia obteve uma taxa de aproximadamente 0,93 falsos positivos por imagem. Com relação ao número de falsos negativos, obteve-se uma taxa de 0,02 falsos negativos por imagem.

Dessa maneira, através da análise dos resultados é possível

concluir que presente metodologia mostrou-se eficaz na detecção de massas em mamografias. Os resultados apresentados permitem inferir que o uso do algoritmo *Growing Neural Gas* na tarefa de segmentação de mamografias permite um bom índice de segmentação correta da massa. Da mesma forma, o uso da função K de Ripley modificada permite a descrição de textura de uma maneira eficiente, contribuindo para o reconhecimento das estruturas segmentadas. Além disso, a classificação MVS apresenta grande poder de generalização, contribuindo para a eficácia da metodologia, durante a fase de classificação dos objetos segmentados.

#### **4.3 Casos de estudo**

O objetivo desta seção é examinar todos os passos da metodologia proposta em testes reais, com o objetivo de detalhar cada etapa e ao mesmo tempo facilitar a compreensão das técnicas utilizadas neste trabalho. Para isso serão examinados três casos. O primeiro corresponde a um caso em que a metodologia obteve êxito total na detecção da massa. Para o segundo caso, a metodologia obteve sucesso na detecção da massa, mas detectou também objetos que não são massas de fato, de acordo com a avaliação realizada por um especialista. O terceiro caso corresponde a um caso em que a metodologia não foi capaz de detectar a massa na imagem. Neste último, serão discutidos os possíveis motivos que levaram à falha do processo de detecção.

### 4.3.1 Primeiro caso

A Figura 4.3 mostra a seqüência de passos utilizados para a detecção correta de uma massa em uma mamografia digitalizada. A mamografia utilizada para testar a metodologia proposta no presente caso de estudo está representada na Figura 4.3-a, sendo indicada em vermelho a localização da massa fornecida pelo DDSM.

O primeiro passo diz respeito ao pré-processamento da imagem. Especificamente, são retirados da imagem todos os objetos que não pertencem à mama em si, tais como os rótulos contendo datas, nomes e tipo de projeção. No presente caso, observa-se que as letras “l”, “m” e “o”, presentes na imagem, indicam o tipo de projeção utilizada (médio lateral oblíqua). A próxima etapa é a segmentação da imagem mamográfica com o intuito de descobrir os candidatos a massa. Para isso, o algoritmo GNG é executado com os parâmetros descritos na Seção 3.3.2. O mesmo algoritmo é executado para 7, 8 e 9 grupos (neurônios). Isso é feito porque muitas vezes o algoritmo não é capaz de separar determinados objetos candidatos a massa utilizando 7 grupos, mas é capaz de separar utilizando 8 ou 9 grupos, ou vice-versa. A Figura 4.3-b mostra o resultado dessa etapa, ou seja, a imagem segmentada utilizando cada número de grupos especificado.

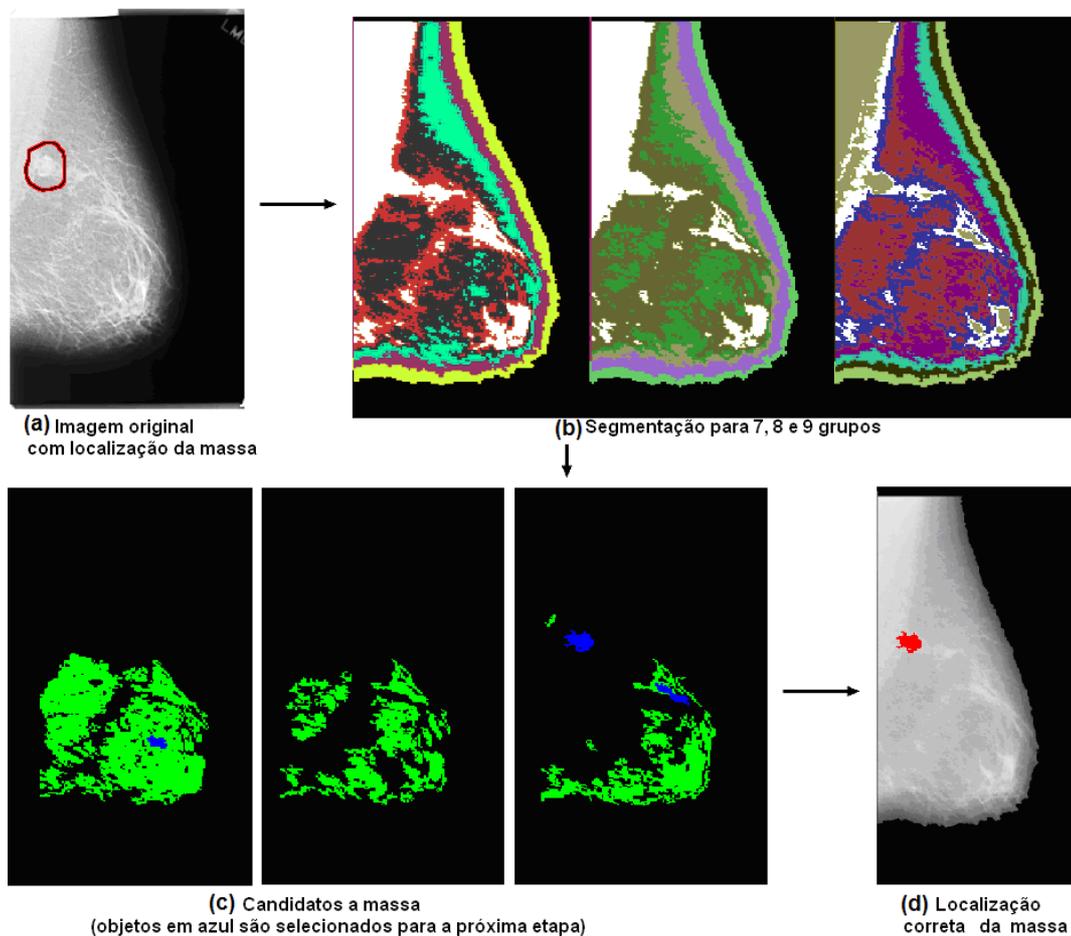


Figura 4.3 – Primeiro caso de estudo

Examinando as imagens segmentadas é possível observar a geometria de determinados objetos os torna descartáveis como possíveis candidatos a massa. Isso porque tais objetos são grandes, alongados ou retorcidos demais. Portanto, a próxima etapa trata da redução dos candidatos a massa, através da análise das características geométricas propostas feita por um classificador MVS. A Figura 4.3-c mostra os objetos classificados como massa (cor azul) e como não-massa (cor verde), de acordo com a sua geometria.

Somente os objetos considerados massa na etapa anterior têm as características de textura extraídas através da função  $K$  de Ripley

modificada, discutida na Seção 2.3.2. Através da classificação MVS, são indicados então os objetos classificados como massa, sendo essa a etapa final do processo de detecção.

A Figura 4.3-d ilustra a detecção bem sucedida, mostrando a massa detectada na cor vermelha.

### **4.3.2 Segundo caso**

A mamografia utilizada para o presente caso é ilustrada na Figura 4.4-a, a qual exhibe a localização da massa fornecida pelo DDSM. O pré-processamento da imagem retira todos os objetos que não pertencem à mama em si. Para a etapa de segmentação novamente o algoritmo é executado para 7, 8 e 9 grupos (Figura 4.4-b).

A próxima etapa trata da redução dos candidatos a massa, através da análise das características geométricas propostas feita por um classificador MVS. Na Figura 4.4-c são observados os objetos classificados como massa e como não-massa, de acordo com a sua geometria. Novamente, objetos azuis correspondem a objetos classificados como massa de acordo com a geometria, e objetos verdes correspondem a objetos classificados como não-massa nessa etapa.

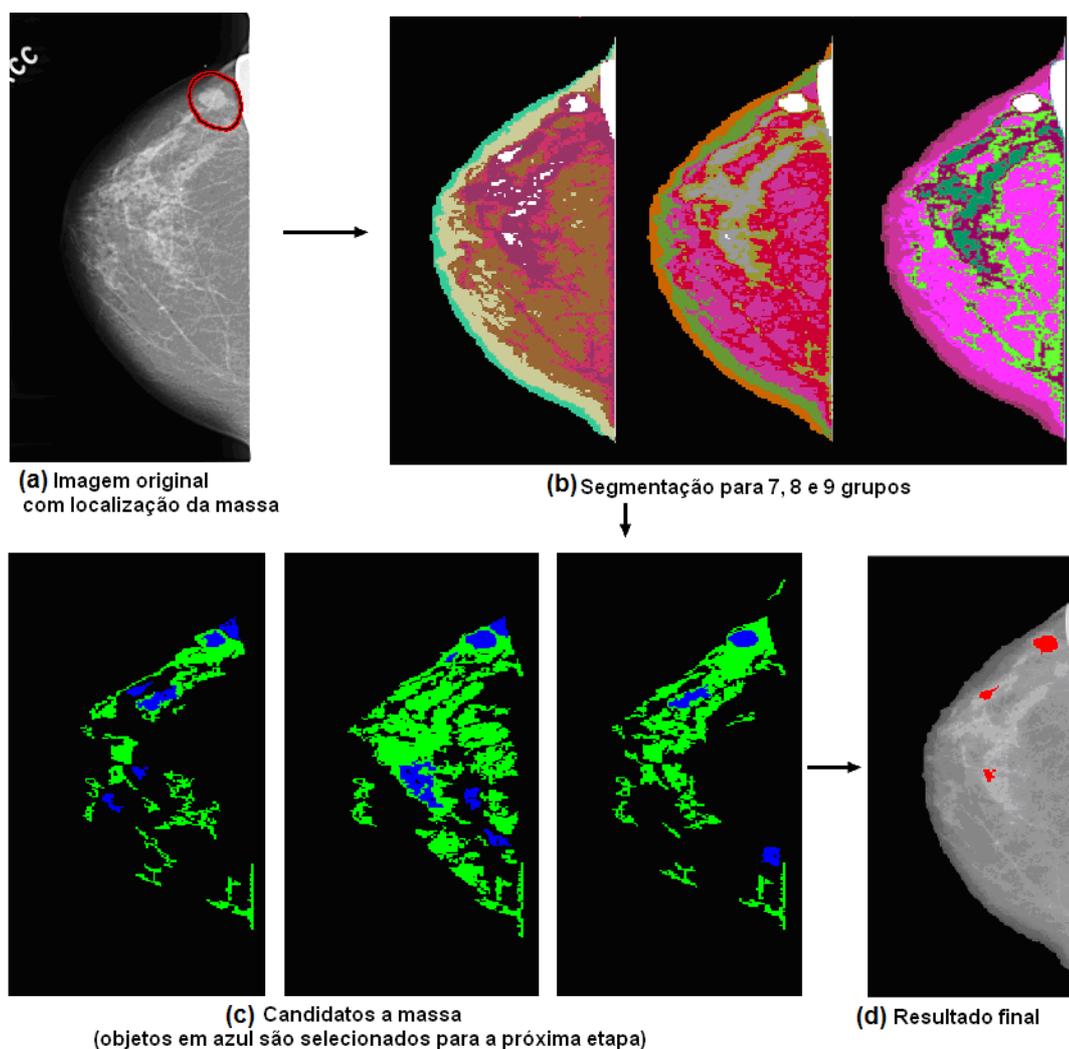


Figura 4.4 – Segundo caso de estudo

O objetos considerados massa na etapa anterior têm as características de textura extraídas através da função  $K$  de Ripley modificada e através da classificação MVS, são indicados quais os objetos que corresponde a massas. A Figura 4.4-d ilustra o resultado final do processo de detecção, mostrando os objetos considerados massa pintados na cor vermelha.

Para o presente caso, observa-se que a metodologia obteve sucesso na detecção da massa, de acordo com a Figura 4.4-d. Entretanto, observa-se também que além da massa correta outro objeto foi indicado

pela metodologia como sendo massa, quando na realidade isso não é verdade. Isso significa que a etapa de redução dos falsos positivos, realizada através da extração da textura via função  $K$  de Ripley modificada e da classificação MVS, não foi capaz de classificar corretamente esse objeto como pertencente à classe não-massa.

#### **4.3.3 Terceiro caso**

No presente caso é examinada uma mamografia que contém uma massa. Entretanto, a metodologia proposta não foi capaz de detectar a mesma. A mamografia utilizada para no presente caso é ilustrada na Figura 4.5-a, a qual exibe a imagem original e a localização da massa fornecida pelo DDSM.

O pré-processamento da imagem retira todos os objetos que não pertencem à mama em si. Para a etapa de segmentação novamente o algoritmo é executado para 7, 8 e 9 grupos. A Figura 4.5-b mostra o resultado dessa etapa. Os candidatos a massa são reduzidos através do uso de medidas geométricas em conjunto com a classificação MVS. A Figura 4.5-c mostra os objetos considerados massa e aqueles considerados não-massa, de acordo com sua geometria.

Os objetos considerados massa de acordo com a geometria são então classificados de acordo com sua textura. Analisando o resultado final da detecção (Figura 4.5-d), observa-se que a metodologia proposta falha em detectar a massa indicada pelo especialista.

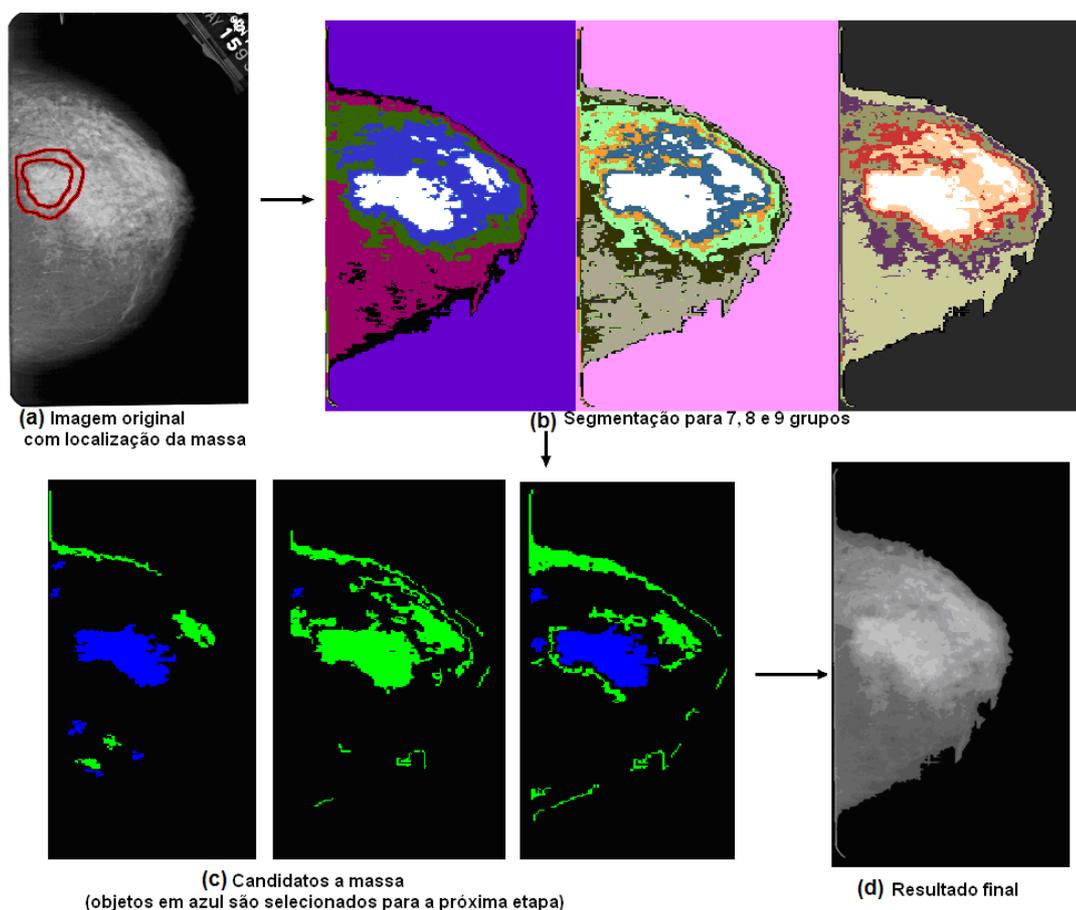


Figura 4.5 – Terceiro caso de estudo

A análise da etapa de segmentação, cujo resultado é visto na Figura 4.5-d indica que o algoritmo GNG não foi capaz de segmentar corretamente a massa especificada, agregando na mesma estrutura a massa e uma grande parte de tecido normal. Uma possível razão para isso deve-se ao fato de que trata-se de uma mama densa, o que faz com que as bordas da massa sejam poucos distinguíveis dos demais tecidos normais. Para casos como esse, a adoção de métodos de realce de contraste poderia fazer com que as bordas da massa tornassem-se mais visíveis, facilitando o trabalho do algoritmo de agrupamento.

## 5 CONCLUSÃO

O número de mortes ocasionadas pelo câncer de mama a cada ano no Brasil ainda permanece bastante elevado. Para que esse número seja reduzido, o câncer precisa ser diagnosticado ainda em seu estado inicial, através de exames capazes de detectá-lo com o máximo de precisão e rapidez possíveis.

O uso de ferramentas computacionais para auxílio à detecção e diagnóstico tem crescido em aceitação nos últimos anos, como uma espécie de “segundo leitor” das imagens médicas. Tais ferramentas têm a capacidade de aumentar o índice de detecção precoce de câncer de mama, conforme visto no Capítulo 1.

O presente trabalho apresentou uma metodologia para a detecção de massas em mamografias digitalizadas, o qual pode ser utilizado no desenvolvimento de uma ferramenta CAD. A metodologia utilizada para tanto subdivide-se em pré-processamento, segmentação através do algoritmo *Growing Neural Gas*, redução dos candidatos a massa, classificação do objeto segmentado em massa ou não-massa com base na função  $K$  de Ripley local. Na etapa de redução dos candidatos a massa são extraídas medidas geométricas do mesmo e essas são submetidas à análise de um classificador MVS.

Dentre as contribuições apresentadas neste trabalho destacam-se o uso do algoritmo *Growing Neural Gas* para a tarefa de segmentação da imagem mamográfica, bem como o uso da função  $K$  de Ripley modificada para a descrição de textura dos objetos segmentados. Os resultados indicam que o uso dessas duas técnicas torna possível a

detecção de massas com razoável eficiência, abrindo o caminho para o desenvolvimento de um sistema CAD capaz de auxiliar profissionais de saúde na árdua tarefa de rastrear a mamografia em busca de anormalidades do tipo massa.

Deste modo, o método apresentado no presente trabalho pode constituir-se em uma ferramenta CAD a ser utilizada para auxiliar a detecção de massas utilizando-se de mamografias digitalizadas, pois foram obtidos resultados que comprovaram a eficiência do mesmo para esse fim.

### **5.1 Trabalhos futuros**

Com a experiência adquirida neste trabalho, são sugeridas as seguintes extensões para esta tese:

- Utilizar a função  $K$  de Ripley para o diagnóstico da massa segmentada em benigno ou maligno;
- Comparar a função  $K$  de Ripley com outros métodos consagrados de extração de textura, tais como as matrizes de co-ocorrência;
- Realizar uma comparação entre o SVM e outros tipos de classificadores, utilizando-se da função  $K$  de Ripley como descritor de textura;
- Comparar a mamografia com outras técnicas de imageamento, tais como a ultra-sonografia e a ressonância magnética, nas tarefas de detecção e diagnóstico de nódulos mamários

- Ampliar a capacidade auto-organizacional do algoritmo *Growing Neural Gas* implementado neste trabalho, criando técnicas para inferir o número ótimo de neurônios necessários para o processo de agrupamento, sem a necessidade de entrar com o número de neurônios como parâmetro de entrada;
- Desenvolver uma interface gráfica que permita ao usuário utilizar a metodologia proposta nesta tese, através do ajuste de parâmetros dos algoritmos propostos bem como o ajuste de brilho e contraste das imagens;
- Investigar o uso do algoritmo *Growing Neural Gas* na segmentação de outros tipos de imagens médicas, tais como a tomografia computadorizada, a ressonância magnética, dentre outros.

## Referências

AZPITARTE, R. L. **Aportaciones al Diagnóstico de Cáncer Asistido por Ordenador**. Tese de Doutorado. Universidad Politécnica de Valencia, 2006.

BELLOTTI, R.; CARLO, F. de; TANGARO, S.; GARGANO, G.; MAGGIPIINTO, G.; CASTELLANO, M.; MASSAFRA, R.; CASCIO, D. **A completely automated CAD system for mass detection in a large mammographic database**. Medical Physics, Volume 33, p. 3066-3075, 2006.

BOYLE, P. **Current situation of screening for cancer**. Annals of Oncology, Switzerland, v. 13, Supl. 4, p. 45-50, 1991.

BRAZ JÚNIOR, G. **Identificação de Massas em Mamografias Usando Textura, Geometria e Algoritmos de Agrupamento e Classificação**. Monografia de conclusão de curso. Universidade Federal do Maranhão, departamento de Ciência da Computação. São Luís, 2006.

CHAN, H.; WEI, J.; SAHINER, B.; RAFFERTY, E.; WU, T.; ROUBIDOX, M. MOORE, R.; KOPANS, D. **Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience**. Radiology, p. 1075 – 1080, 2005.

CHANG, Chih-Chung; LIN, Chih-Jen. **LIBSVM: a Library for Support Vector Machines**. Disponível em <<http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>>. Acesso em 08 out. 2007.

CHAVES, A. C. F. **Extração de regras fuzzy para Máquinas de vetor de Suporte (SVM) para classificação em múltiplas classes**. Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro, 2006.

CRISTIANINI, N.; SHAWE, J. **An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods**. Press Syndicate of the University of Cambridge, 2000.

FIGUEIREDO, A. M. **Implementação do algoritmo de clusterização Growing Neural Gas**. Projeto final de Programação. Pontifícia Universidade Católica do Rio de Janeiro – PUC-RJ. Rio de Janeiro, 2006.

FRITZKE, B; **A growing neural gas network learns topologies**. Em G. Tesauro, D. S. Touretzky, e T. K. Leen, editores. *Advances in Neural Information Processing Systems 7*, p. 625--632. MIT Press, Cambridge MA, 1995.

GONZALEZ, R.C.; WOODS, R. E. **Digital image processing**. Addison Wesley, 1992.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. **Texture Features form Image Classification**. *IEEE Transactions on Systems, man and Cybernetics*, v. 3, n. 6, p. 610-621, 1973.

HAYKIN S. **Redes Neurais: Princípios e prática**. 2ed. Porto Alegre: Bookman, 2001.

HEATH, M; BOWYER, K.W; KOPANS, D. **Current status of the Digital Database for Screening Mammography**. *Digital Mammography*, p. 457-460. Kluwer Academic Publishers, 1998

Instituto Nacional do Câncer - INCA. **Atlas de Mortalidade por Câncer no Brasil 1979-1999**. Rio de Janeiro: INCA, 2002. Disponível em: <[http://www.inca.gov.br/atlas/docs/distrib\\_prop\\_MA.pdf](http://www.inca.gov.br/atlas/docs/distrib_prop_MA.pdf)>. Acesso em 10 dez. 2005.

Instituto Nacional do Câncer – INCA. **Câncer no Brasil: dados dos registros de base populacional, volume 3**. Rio de Janeiro: INCA, 2003. Disponível em: <<http://www.inca.gov.br/regpop/2003/versaofinal.pdf>>. Acesso em 15 set. 2006.

Instituto Nacional do Câncer - INCA. **Estimativa 2006: Incidência de Câncer no Brasil**. Rio de Janeiro: INCA, 2005. Disponível em <<http://www.inca.gov.br/estimativa/2006/versaofinal.pdf>>. Acesso em 19 fev. 2006.

Instituto Nacional do Câncer - INCA. **Estimativas de incidência e mortalidade por câncer no Brasil**. Rio de Janeiro: INCA, 2005.

Institut für Neuroinformatik. **Growing Self-Organizing Networks**. Disponível em <<http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/DemoGNG/GNG.html#top>>. Acesso em 19 jun. 2007.

KOPANS, D. B., **Imagem da Mama**. Porto Alegre: MEDSI, 2000.

MARSHALL, D. **Statistical Region Description**. Disponível em <[http://www.cm.cf.ac.uk/Dave/Vision\\_lecture/node36.html](http://www.cm.cf.ac.uk/Dave/Vision_lecture/node36.html)>. Acesso em 15 out. 2007.

MARTINEZ, T.; SCHULTEN, K. **A neural gas network learns topologies**. Artificial Neural Networks Elseviers Science Publishers B.V, 1991.

MARTINS, L. de O.; BRAZ JUNIOR, G. S SILVA, E. C.; S SILVA, A.C.; PAIVA, A.C. **Classification of Breast Tissues in Mammogram Images Using Ripley's K Function and Support Vector Machine**. [ICIAR 2007](#), p. 899-910, 2007a.

MARTINS, L. de O.; SILVA, E. C.; SILVA, A.C.; PAIVA, A.C; GATTASS, M. **Classification of Breast Masses in Mammogram Images Using Ripley's K Function and Support Vector Machine**. [MLDM 2007](#): 784-794, 2007b.

OFFICE FOR NATIONAL STATISTICS - ONS. **Long-term Breast Cancer Survival, England and Wales, up to 2003**. Disponível em <<http://www.statistics.gov.uk/statbase/Product.asp?vlnk=14172&More=Y>> Acesso em 31 out. 2007.

PADWAL, M. **Elements of breast imaging basics**. Disponível em <[http://www.gehealthcare.com/usen/ultrasound/education/products/cme\\_breast.html](http://www.gehealthcare.com/usen/ultrasound/education/products/cme_breast.html)>. Acesso em 03 out. 2007.

PAIVA, J. A. C.; RODRÍGUEZ, A.; CORREIA, V. R. M. **Métodos Computacionais para Analisar Padrões de Pontos Espaciais**. Instituto Nacional de Pesquisas Espaciais, 1999. Disponível em <[http://www.dpi.inpe.br/geopro/trabalhos/gisbrasil99/estat\\_pontos/](http://www.dpi.inpe.br/geopro/trabalhos/gisbrasil99/estat_pontos/)>. Acesso em 02 set. 2007.

RANGAYYAN, R.M.; EL-FARAMAWY, N.M.; DESAUTELS, J.E.L.; ALIM, O.A. **Measures of Acutance and Shape for Classification of Breast Tumors**, Associate Member, IEEE, 1997.

REZAI-RAD, G.; JAMARANI, S. **Detecting Microcalcification Clusters in Digital Mammograms Using Combination of Wavelet and Neural Network**, *cgiv*, p. 197-201, International Conference on Computer

Graphics, Imaging and Visualization (CGIV'05), 2005.

RIPLEY, B. D. **Modelling spatial patterns**. J. Roy. Statist. Soc. B 39, p. 172–212, 1977.

SAMPAT, P; MARKEY, Mia K; BOVIK, Alan C. **Computer-aided detection and diagnosis in mammography**. **Computer-Aided Detection and Diagnosis in Mammography**. *Handbook of Image and Video Processing*, 2ed, 2005

SCHOUTEN, T. **Image Processing**. Radboud University, Department of Computer Science. Disponível em <<http://www.cs.ru.nl/~ths/rt2/col/h9/9gebiedENG.html>>. Acesso em 23 out. 2007.

SILVA, A. G. **O Câncer de Mama**. Universidade do Estado de Santa Catarina. Disponível em <<http://www.joinville.udesc.br/processamentodeimagens/oCanc/oCanc.html>>. Acesso em 21 out. 2007.

SILVA, D. F. **Câncer de mama em mulheres no Maranhão: estudo de sobrevivida no centro de assistência de Alta Complexidade em oncologia (CACON) em São Luís**. Dissertação de Mestrado (Mestrado em Ciências da Saúde). Universidade Federal do Maranhão, 2006.

SILVA, E. C.; SILVA, A.C.; PAIVA, A.C; NUNES, R. A.; GATTASS, M. **Diagnosis of Lung Cancer using Ripley's K function applied to concentric spherical layers from pulmonary nodules CT images**. Submetido à *Pattern Recognition*, 2006.

SILVA, J. M. S.; SOUSA, L.A. **Implementação Paralela de um Algoritmo de Triangulação de Delaunay**. Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento – INESC-ID. Lisboa, 2001.

Sociedade Brasileira de Mastologia - SBM. **Campanha Nacional de Câncer de Mama 2007**. Disponível em <<http://www.revistahands.com.br/noticia.php?noticia=225>>. Acesso em 29 out. 2007.

SOUSA, J. R. F. S. ; SILVA, A. C. ; PAIVA, A. C. **Lung Structures Classification Using 3D Geometric Measurements and SVM**. In: 12th Iberoamerican Congress on Pattern Recognition - CIARP 2007, 2007,

Valparaíso. Lecture Notes Computer Science - LNCS. Berlin : Springer-Verlag, 2007. v. 4756. p. 783-792.

TSCHUMPERLÉ, D. **The Cimg Library Reference Manual**. Disponível em [http://cimg.sourceforge.net/Cimg\\_reference.pdf](http://cimg.sourceforge.net/Cimg_reference.pdf). Acesso em 05 nov. 2007.

University of Guelph. **Shape Analysis & Measurement**, CIS\*6320, Lecture Notes, 2001.

URBAN, D. L. **Spatial analysis in ecology - point pattern analysis**. Disponível em <http://www.nicholasdukeedu/lel/env352/ripley.pdf>, 2003. Acesso em 15 jun. 2007.

VALLE, J. C. do , **Câncer de Mama Locorregional Avançado**. Rio de Janeiro: MEDSI, 1999.

VAPNICK, V. **Statistical Learning Theory**. J.Wiley and Sons Inc. Nova York, 1998.