Suporte a sistemas de auxílio ao diagnóstico e de recuperação de imagens por conteúdo usando mineração de regras de associação

Marcela Xavier Ribeiro

Suporte a sistemas de auxílio ao diagnóstico e de recuperação de imagens por conteúdo usando mineração de regras de associação[1]

Marcela Xavier Ribeiro

Orientadora: Profa. Dra. Agma Juci Machado Traina

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Novembro de 2008

[1] Apoio financeiro FAPESP (processo 04/02215-5)

Ao meu marido Sérgio, meu companheiro,
a quem muito tenho a agradecer.

Agradecimentos

À Deus, por sempre estar ao meu lado.

À minha orientadora Agma, por me direcionar, ensinar, incentivar e, principalmente, por ser um exemplo a quem sempre seguirei.

Aos meus pais Pedro e Izabel, minha irmã Renata, meu tio Sale, meu sobrinho Jorge, meu cunhado Nestor e meu amigo Pelé, que são minha turma de apoio em Dourado, por sempre fazerem o possível e o impossível por mim.

Ao prof. Caetano Traina, pelo apoio e por tantas excelentes idéias.

A Nancy e Miriam, por me acolher com tanto afeto durante minha estadia nos Estados Unidos.

Aos meus amigos - Joselene, Joaquim, Elaine, André, Carol, Pedro, Natália, Mônica, Luciana, Caio, Júnior, Ana Paula, Adriano, Marcelo, Camila, Humberto, Renato, Robson, Daniel, Sérgio e Willian e demais membros do GBDI - por sempre me ajudarem muito.

Aos professores Paulo Marques, Christos Faloutsos e Eric Xing, por me auxiliarem neste trabalho de doutorado.

Aos meus sogros Sérgio e Marizilda, por tudo que fizeram por mim.

À FAPESP pelo apoio financeiro.

Resumo

Neste trabalho, a mineração de regras de associação é utilizada para dar suporte a dois tipos de sistemas médicos: os sistemas de busca por conteúdo em imagens (*Content-based Image Retrieval* - CBIR) e os sistemas de auxílio ao diagnóstico (*Computer Aided Diagnosis* - CAD). Na busca por conteúdo, regras de associação são empregadas para reduzir a dimensionalidade dos vetores de características que representam as imagens e para diminuir o "gap semântico", que existe entre as características de baixo nível das imagens e seu significado semântico.

O algoritmo StARMiner (**St**atistical **A**ssociation **R**ule **Miner**) foi desenvolvido para associar características de baixo nível das imagens com o seu significado semântico, sendo também utilizado para realizar seleção de características em bases de imagens médicas, melhorando a precisão dos sistemas CBIR. Para dar suporte aos sistemas CAD, o método IDEA (***I**mage **D**iagnosis **E**nhancement through **A**ssociation rules*) foi desenvolvido. Nesse método regras de associação são empregadas para sugerir uma segunda opinião ou diagnóstico preliminar de uma nova imagem para o radiologista. A segunda opinião automaticamente gerada pelo método pode acelerar o processo de diagnóstico de uma imagem ou reforçar uma hipótese, trazendo ao especialista médico um apoio estatístico da situação sendo analisada. Dois novos algoritmos foram propostos: um para pré-processar as características de baixo nível das imagens médicas e, o outro, para propor diagnósticos baseados em regras de associação. Vários experimentos foram realizados para validar os métodos desenvolvidos. Os experimentos realizados indicam que o uso de regras de associação pode contribuir para melhorar a busca por conteúdo e o diagnóstico de imagens médicas, consistindo numa poderosa ferramenta para descoberta de padrões em sistemas médicos.

Abstract

In this work we take advantage of association rule mining to support two types of medical systems: the Content-based Image Retrieval (CBIR) and the Computer-Aided Diagnosis (CAD) systems. For content-based retrieval, association rules are employed to reduce the dimensionality of the feature vectors that represent the images and to diminish the semantic gap that exists between low-level features and its high-level semantical meaning.

*The StARMiner (**St**atistical **A**ssociation **R**ule **Miner**) algorithm was developed to associate low-level features with their semantical meaning. StARMiner is also employed to perform feature selection in medical image datasets, improving the precision of CBIR systems. To improve CAD systems, we developed the IDEA (**I**mage **D**iagnosis **E**nhancement through **A**ssociation rules) method. Association rules are employed to suggest a second opinion to the radiologist or a preliminary diagnosis of a new image. A second opinion automatically obtained can accelerate the process of diagnosing or strengthen a hypothesis, giving to the physician a statistical support to the decision making process. Two new algorithms are developed to support the IDEA method: to pre-process low-level features and to propose a diagnosis based on association rules. We performed several experiments to validate the developed methods. The results indicate that association rules can be successfully applied to improve CBIR and CAD systems, empowering the arsenal of techniques to support medical image analysis in medical systems.*

Sumário

Lista de Figuras

Lista de Tabelas

Glossário de Termos

***k*NN**	*k-nearest neighbor* (k-vizinhos mais próximos), 43
a/p	anterior/posterior, 157
CAD	*Computer-aided Diagnosis* (Diagnóstico Auxiliado por Computador), 2
CBIR	*Content-based Image Retrieval* (Recuperação de Imagens Baseada em conteúdo), 3
CCIFM	Centro de Ciências de Imagens e Física Médica, 4
Classe Majoritária	a classe mais freqüente nas instâncias pertencentes a um intervalo, 57
CR	*Computed Radiography* (Radiografia computadorizada), 33
CT	*Computed Tomography* (Tomografia Computadorizada), 33
d/v	dorsal/ventral, 157
DICOM	*Digital Imaging and Communications in Medicine*, é um protocolo de comunicação e formato de imagem médica, 33
DM	*Data Mining* (Mineração de Dados), 2
DR	*Direct Digital Radiography* (Radiografia Digital Direta), 33
DTM	*Decision Tree Method* (Método baseado em Árvore de Decisão), 61
EM/MPM	***E**xpectation **M**aximization / **M**aximization of the **P**osterior **M**arginals*, 18
GBDI	Grupo de Base de Dados e Imagens do ICMC (USP - São Carlos), 4
ICMC	Instituto de Ciências Matemáticas e de Computação, 4
Inconsistência	ocorre quando uma instância de classe diferente da classe majoritária é alocada em um intervalo, 57

KDD *Knowledge Discovery in Databases* (Processo de Descoberta de Conhecimento em Base de Dados), 52

MAE Método de Acesso Espacial, 42
MAM Método de Acesso Métrico, 42
MDLP *Minimum Description Length Principle*, princípio do comprimento mínimo da descrição, 58
Mineração de Imagens refere-se à extração automática ou semi-automática de padrões de imagens, 51
Mineração Visual consiste em visualizar grandes conjuntos de dados e seus padrões com o intuito de verificar tendências nos mesmos, 51
MR *Magnetic Resonance* (Ressonância Magnética), 33

P&R *Precision vs. Recall* (Precisão versus Revocação), 47
PACS *Picture Archiving and Communication Systems* (Sistemas de Arquivamento e Comunicação de Imagens), 2
PCA *Principal Components Analysis*, Análise de Componentes Principais, 60
Ponto de corte limite de um intervalo de valores reais, 57

Regra forte regra que satisfaz o suporte mínimo e a confiança mínima, 65
RM Ressonância Magnética, 90
ROC *Receiver Operating Characteristic*, 34

SGBD Sistemas Gerenciadores de Banco de Dados, 4
SIH Sistemas de Informação Hospitalar, 33
SIR Sistemas de Informação Radiológica, 33

USP Universidade de São Paulo, 4

validação *k-fold* o conjunto de dados é dividido em k conjuntos de tamanhos iguais para a validação do algoritmo de mineração. O algoritmo de mineração é executado k vezes. Para cada execução, $k-1$ conjuntos de dados são utilizados para treinar e 1 conjunto é utilizado para teste. O conjunto de teste varia em cada uma das execuções, 111

Capítulo 1

Introdução

1.1 Considerações Iniciais

O aperfeiçoamento dos equipamentos eletrônicos e dos sistemas computacionais contribui para o desenvolvimento de muitas áreas de pesquisa, e a medicina é uma das áreas que mais tem se beneficiado desse aperfeiçoamento. O volume de dados médicos armazenados, que incluem exames, diagnósticos e procedimentos de tratamento, tem uma tendência de crescimento exponencial. Conforme as leis vigentes em nosso país, esses dados devem ser guardados por no mínimo 20 anos (resolução do Conselho Federal de Medicina, número 1.821/2007, www.conarq.arquivonacional.gov.br). Esse grande volume de dados históricos é uma valiosa fonte de conhecimento, que pode ser usada para o auxílio ao diagnóstico médico e para o ensino da medicina. No entanto, em virtude da complexidade da análise e tratamento dos dados que incluem imagens, os profissionais da área de saúde ainda não se beneficiam de grande parte dessa fonte de conhecimento. Por exemplo, as técnicas existentes para a recuperação de imagens dificilmente permitem que sejam encontradas imagens de exames antigos com o mesmo tipo de anomalia mostrado em uma imagem recém-obtida. Esse fato ocorre em virtude de haver um grande número de características extraídas das imagens que podem ser usadas para sua busca, mas são desconhecidas quais delas são as mais relevantes para cada tipo de aplicação. Além disso, o uso de um grande número de características pode levar ao problema da "maldição da alta dimensionalidade" [Jeong et al., 2007], que degrada a precisão e o tempo de busca. Devido a tais desafios, as técnicas de recuperação de imagens por conteúdo têm sido bastante pesquisadas nos últimos anos.

Pesquisadores das comunidades de banco de dados, inteligência artificial, estatística, aprendizado de máquina, visualização e computação paralela, entre outras, têm trabalhado de maneira integrada para desenvolver técnicas eficientes para manipular e compreender o inter-relacionamento de grandes conjuntos de dados. Desse esforço conjunto surgiu a área de mi-

neração de dados (*data mining* - DM) [Fayyad et al., 1996] A mineração de dados é definida como "um processo não trivial de extração de informações implícitas e previamente desconhecidas que são potencialmente úteis para a compreensão efetiva do conjunto de dados em análise" [Frawley et al., 1991]. Considerando imagens, deve-se ressaltar que a mineração desse tipo de dados é uma tarefa muito importante para a área médica, onde para a obtenção de diagnósticos precisos baseados em imagens, é necessária uma análise complexa das imagens envolvidas, comparando um grande número de características. Tal situação indica a demanda por técnicas de mineração de imagens.

Existem na literatura várias tarefas de mineração que permitem extrair conhecimento sobre os dados. Neste trabalho a tarefa de associação é explorada com maior ênfase. A mineração de regras de associação visa extrair regras que relacionam itens que freqüentemente ocorrem juntos nas bases de dados analisadas. Um exemplo de padrão que pode ser obtido analisando imagens médicas é o conhecimento de que um determinado tipo de tumor está associado a uma característica específica do tecido, posição ou relacionamento com outros objetos da imagem. Além disso, associações refletem como os seres humanos adquirem novos conhecimentos e memorizam, tendo uma vasta aplicabilidade em várias áreas de conhecimento.

1.2 Motivação

Atualmente, muitos sistemas computacionais precisam analisar dados complexos, tais como imagens, vídeos, séries temporais, impressões digitais e seqüências de DNA. Em se tratando de imagens, que é um dos dados complexos mais estudados, e mais especificamente de imagens médicas, atualmente dois tipos de sistemas são amplamente utilizados: os Sistemas de Arquivamento e Comunicação de Imagens (*Picture Archiving and Communication Systems* - PACS) e os sistemas de Diagnóstico Auxiliado por Computador (*Computer-aided Diagnosis* - CAD).

Pesquisadores de técnicas para uso em sistemas CAD têm pesquisado algoritmos eficientes e precisos para classificar corretamente imagens médicas em categorias relevantes, com a intensão de auxiliar na tarefa de diagnóstico de imagens médicas. De fato, o desenvolvimento dos sistemas PACS ampliou o uso efetivo de imagens no diagnóstico médico e também no ensino da medicina [Muller et al., 2004]. Entretanto, para que sejam efetivamente úteis, o processamento e a recuperação das imagens nos sistemas PACS e CAD devem ser rápidos e consistentes com o julgamento dos especialistas.

O volume de imagens gerado em exames médicos cresce exponencialmente, demandando métodos eficientes e efetivos de análise e recuperação de imagens. Isso demanda que técnicas eficazes de armazenamento e recuperação de imagens sejam desenvolvidas para que o acesso a dados históricos seja possível. Assim, as técnicas de busca por conteúdo (*Content-based Image*

Retrieval - CBIR) têm sido intensamente investigadas nos últimos anos [Liu et al., 2007] e os sistemas CBIR vêm sendo incorporados aos sistemas PACS.

As técnicas de CBIR tipicamente utilizam características automaticamente ou semi-automaticamente extraídas das imagens através de algoritmos de processamento de imagens. Essas características são agrupadas em vetores de características, que são armazenados e organizados em uma estrutura de indexação para possibilitar um acesso rápido e eficiente às imagens. Geralmente, os sistemas CBIR utilizam características intrínsecas (de baixo nível) das imagens, tais como cor, forma e textura [Kinoshita et al., 2007] levando a vetores com centenas a milhares de características. No entanto, o uso de um grande número de características representa um problema. A medida que o número de características aumenta, o processo de armazenamento, indexação, recuperação e comparação das imagens se torna cada vez mais lento [Egecioglu et al., 2004] e, em muitas situações, muitas características são correlacionadas, trazendo informações redundantes, o que pode deteriorar a habilidade do sistema distinguir corretamente as imagens. O uso de um grande número de características fazem os sistemas CBIR sofrerem com o problema conhecido como a "maldição da alta dimensionalidade" [Aggarwal, 2001]. Beyer em [Beyer et al., 1999] provou que, a medida que o número de características aumenta, a significância de cada característica tende a diminuir. Assim, é importante manter o tamanho do vetor de características o menor possível, estabelecendo uma relação de compensação entre o tamanho do vetor de características e sua capacidade de representação da imagem. Por isso, é importante o desenvolvimento de técnicas de mineração capazes de identificar automaticamente as características mais relevantes para discriminar as imagens médicas. Um exemplo significativo do uso de características para representar imagens é o uso de características de forma para representar lesões de mama. Inicialmente o radiologista classifica as imagens baseado na forma da lesão. Tumores malignos geralmente infiltram nos tecidos adjacentes, resultando em um contorno irregular, enquanto tumores benignos têm uma borda bem delimitada, conforme é ilustrado na Figura 1.1.

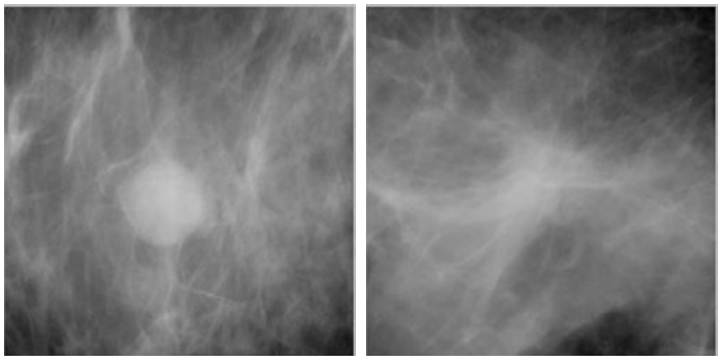


Figura 1.1: Exemplo de tumor benigno (esquerda) e maligno (direita).

Além dos sistemas de CBIR, os sistemas CAD tem auxiliado os radiologistas no diagnóstico

de imagens médicas. Pesquisas recentes mostram que o uso de sistemas CAD melhora significativamente o desempenho dos radiologistas em detectar anomalias corretamente. Os sistemas CAD têm auxiliado no diagnóstico e na prevenção de anomalias na mama, pulmão, intestino e cérebro, entre outras partes do corpo. Estudos recentes mostram que radiologistas falham em identificar manualmente 30% dos casos positivos de câncer de pulmão [Doi, 2005]. Em [Quek et al., 2003] foi apresentado um estudo que revelou um aumento estatisticamente significativo (13%) no desempenho de estudantes de medicina na detecção de anomalias em mamografias, ao usar um sistema CAD. Uma série de estudos [Doi, 2007] [Armato III et al., 2001] [Buhmann et al., 2007] permitiram a comparação da precisão de diagnósticos realizados por radiologistas sem auxílio de um sistema CAD e por radiologistas auxiliados por um sistema CAD. Os resultados desses estudos mostraram que o uso de sistemas CAD melhora estatisticamente a precisão dos radiologistas na distinção de nódulos, microcalcificações e assimetrias em mamografias, na análise do tamanho do coração, e na detecção de embolia pulmonar, entre outros. Esses estudos ressaltam a importância do desenvolvimento de ferramentas e técnicas computacionais para auxílio ao diagnóstico de imagens médicas.

Assim, se faz necessário o desenvolvimento de novas técnicas de mineração de imagens para suportar o auxílio ao diagnóstico de imagens e exames médicos. Neste sentido, o Grupo de Base de Dados e Imagens (GBDI) do Instituto de Ciências Matemáticas e de Computação (ICMC) - USP tem desenvolvido, desde 1997 [Traina Jr. et al., 1997], trabalhos relativos a representação de imagens através de vetores de características; possibilitando o armazenamento, indexação e busca por conteúdo de imagens em Sistemas Gerenciadores de Banco de Dados (SGBDs). As pesquisas relacionadas a imagens médicas têm sido desenvolvidas em conjunto com o Centro de Ciências de Imagens e Física Médica (CCIFM) do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto - USP (HCFMRP). Foi justamente este convênio entre o GBDI e o HCFMRP que permitiu o desenvolvimento desta tese que aplica técnicas de mineração de dados para o auxílio ao diagnóstico e aperfeiçoamento da busca por conteúdo em imagens médicas.

1.3 Objetivos

O objetivo desta tese foi definir e desenvolver *métodos* para a construção de sistemas de apoio ao diagnóstico (CAD) e para apoiar sistemas de recuperação de imagens por conteúdo (CBIR) utilizando regras de associação mineradas de imagens e exames médicos, baseadas em características extraídas das próprias imagens e de seus laudos. Com base nos *métodos* propostos foi definido um protótipo de sistema para auxílio ao diagnóstico, usando algoritmos de mineração de imagens e regras de associação.

Na maioria das vezes, é necessário o tratamento dos dados de baixo nível para gerar in-

tervalos compactos e semanticamente mais significativos para minerar regras de associação, permitindo relacionar as informações de baixo nível automaticamente extraídas das imagens com o conhecimento de alto nível do especialista [Abraham et al., 2006]. Assim, esta tese também objetivou o desenvolvimento de técnicas de pré-processamento de dados para a tarefa de associação.

O grau de complexidade do processo de obtenção de regras de associação é proporcional ao número de atributos a serem relacionados. Como o número de elementos (dimensão) dos vetores de características das imagens é elevado (podendo atingir centenas de elementos), é necessário encontrar as características mais relevantes para a mineração das mesmas. Um dos objetivos desta tese foi também utilizar regras de associação para sintetizar o conjunto de características das imagens.

Nesta tese, buscou-se modelar e desenvolver métodos para a identificação de regras de associação em imagens que possam efetuar a tarefa de modo mais eficiente do que o atual estado-da-arte, utilizando técnicas de seleção de atributos e processamento incremental. Desse modo, buscou-se contornar a "maldição de alta dimensionalidade" [Malcok et al., 2006], que é um dos principais problemas tratados pelos pesquisadores da área de indexação e recuperação de dados complexos.

1.4 Desafios

Um desafio para a área de mineração de imagens médicas é o desenvolvimento de técnicas para reduzir o "gap semântico" entre as características de baixo nível da imagem (representação numérica) e a interpretação humana (descrição semântica). O uso da mineração de regras de associação pode ajudar a efetivamente reduzir o "gap semântico" de uma maneira automática, onde padrões relacionando conteúdo semântico com a representação de baixo nível das imagens podem ser encontrados. O problema da "maldição da alta dimensionalidade" é outro desafio a ser vencido ao se trabalhar com busca e análise automática de imagens. O emprego de técnicas adequadas de seleção de características é uma das maneiras de vencer esse problema.

As técnicas de auxílio ao diagnóstico devem utilizar o conhecimento do especialista sobre exames e imagens anteriormente analisadas para auxiliar no diagnóstico de novas imagens. Um grande desafio desta tese foi desenvolver um método eficiente, eficaz e confiável de auxílio à tomada de decisão que forneça múltiplas hipóteses de diagnósticos para os especialistas trabalharem, indicando quais hipóteses são mais prováveis. No entanto, para vencer esse grande desafio, outros problemas devem ser transpostos, e duas questões devem ser respondidas:

- Como cruzar informações de baixo nível (características) das imagens com informações de alto nível (laudo) e utilizar esse cruzamento para sugerir novos diagnósticos?

- Como gerar hipóteses de diagnósticos e classificá-las de acordo com sua probabilidade de serem verdadeiras?

A proposta desta tese foi utilizar regras de associação, que é uma abordagem diferente da tradicional classificação utilizada em sistemas CAD. Assim, o uso de regras de associação e sua aplicação para auxílio ao diagnóstico de imagens foi um desafio por si só.

1.5 Resultados Obtidos

Esta tese traz contribuições para as áreas de mineração de dados, mineração de imagens, CAD e CBIR. As principais contribuições são:

- Um novo algoritmo de mineração de regras de associação estatísticas (StARMiner - *Statistical Association Rule Miner*). O algoritmo StARMiner usa medidas estatísticas que descrevem o comportamento das características visuais das imagens considerando diferentes categorias de imagens. O algoritmo identifica através das regras mineradas, as características mais importantes para discriminar as imagens em categorias (por exemplo, tumor benigno e tumor maligno). O StARMiner encontra regras de associação que relacionam dados categóricos (nominais) com dados quantitativos (contínuos), realizando também a tarefa de seleção de características. Esse algoritmo foi aplicado com sucesso na busca por conteúdo usando várias bases de imagens médicas, conforme apresentado no capítulo 5, promovendo redução de dimensionalidade, eliminando características ruidosas e redundantes do vetor de características, e também promovendo um aumento na precisão das buscas [Ribeiro et al., 2005a] [Ribeiro et al., 2006d] [Ribeiro & Traina, 2005]. O algoritmo StARMiner também foi aplicado junto com o algoritmo FDASE para encontrar um conjunto mínimo de atributos para representar imagens de mama, alcançando uma redução ainda maior do vetor de características e mantendo os valores de precisão [Felipe et al., 2006] [Ribeiro et al., 2005b]. As regras geradas pelo StARMiner também foram aplicadas para ponderar os vetores de características na busca por conteúdo, utilizando várias funções de distância, onde ganhos expressivos na precisão das consultas, com significativa redução de dimensionalidade foram alcançados [Bugatti et al., 2008]. Um resultado secundário desta tese foi a utilização do algoritmo StARMiner com técnicas de *realimentação de relevância* para maximizar a precisão das consultas por conteúdo [Ribeiro et al., 2006c] [Ribeiro et al., 2006b] [Ribeiro et al., 2006a].

- O desenvolvimento do algoritmo Omega [Ribeiro et al., 2008d] [Ribeiro et al., 2008a], para o pré-processamento de dados para a tarefa de mineração, realizando duas operações simultaneamente: discretização e seleção de características. O desempenho do algoritmo

Omega nas tarefas de discretização e seleção de atributos foi comparado com outros algoritmos da literatura, atingindo resultados bastante promissores.

- O desenvolvimento do algoritmo ACE, um classificador especial, que retorna um conjunto de palavras-chave para compor uma sugestão de diagnóstico para uma imagem. O ACE usa uma medida de *convicção* (confiabilidade) para determinar a probabilidade da hipótese sugerida pertencer ao diagnóstico final dado pelo radiologista.

- O desenvolvimento do método IDEA, para auxiliar no diagnóstico de imagens médicas. O método IDEA se baseia em regras de associação para gerar automaticamente sugestões de diagnóstico para imagens médicas. Essas sugestões são usadas como uma segunda opinião pelo radiologista [Ribeiro et al., 2008b].

- O início do projeto de Mineração de Embriões de *Drosophila* durante a realização de um estágio pela aluna na Universidade *Carnegie Mellon* em Pittsburgh - EUA. Esse projeto é descrito no Apêndice A desta tese.

1.6 Organização do Trabalho

Esta tese está organizada em 8 capítulos, a saber:

- O capítulo 1 discutiu a introdução, a motivação e os objetivos desta tese;

- O capítulo 2 apresenta conceitos relacionados com o processamento de imagens. Esses conceitos são utilizados para processar as imagens, segmentá-las e extrair características das mesmas, representando-as em vetores de características. Os vetores de características são utilizados como entrada para os métodos de mineração de imagens;

- No capítulo 3 são apresentados os sistemas de suporte à área médica. A pesquisa desenvolvida nesta tese visou desenvolver ferramentas computacionais para o suporte desses sistemas;

- O capítulo 4 revê conceitos a respeito de mineração de dados e mineração de imagens, discutindo as principais tarefas de mineração de dados e imagens, enfatizado a mineração de regras de associação, que é o foco principal desta tese;

- O capítulo 5 descreve o algoritmo StARMiner desenvolvido durante o trabalho desta tese. O algoritmo StARminer é um algoritmo de mineração de regras de associação estatísticas, cujo objetivo principal é minerar regras relacionando características e categorias de imagens médicas;

- O capítulo 6 discute o algoritmo Omega, desenvolvido durante o trabalho desta tese, que é um algoritmo de discretização e seleção de atributo, cujo principal objetivo é pré-processar as características das imagens médicas para a tarefa de mineração de regras de associação;
- O capítulo 7 descreve o método IDEA de auxílio ao diagnóstico baseado em regras de associação, que foi o principal resultado obtido durante este trabalho de tese;
- Finalmente, o capítulo 8 apresenta as conclusões, as principais contribuições desta tese e as possibilidades de pesquisas futuras.

Parte I

Conceitos e Embasamento Teórico

Capítulo 2

Representação de Imagens

2.1 Considerações Iniciais

O enfoque principal deste trabalho é a descoberta de conhecimento em imagens médicas. Um problema de se trabalhar com imagens é que, em geral, a matriz de pixels não agrega nenhum significado que corresponda diretamente à nossa percepção visual.

Um dos principais desafios ao se trabalhar com imagens é encontrar a melhor representação numérica que sintetize a essência da imagem através de um vetor de características. Neste caso, uma imagem pode ser considerada um ponto no espaço definido pelas suas características.

Antes do processo de descoberta de conhecimento, as imagens necessitam ser "traduzidas" para uma representação sintética e bem definida em termos matemáticos, para que algoritmos de mineração possam ser aplicados para a extração de padrões. Essa mesma representação de imagens também é utilizada pelos sistemas CBIR, onde as características visuais extraídas das informações dos pixels das imagens são usadas para representá-las. A figura 2.1 ilustra como o processo de obtenção da representação da imagem é a primeira etapa, tanto no processo de mineração de imagens, quando nos sistemas CBIR.

O processo de obtenção de uma representação automática das imagens envolve: (a) processamento e segmentação de imagens; (b) extração de característica de imagens. Sendo que a etapa (a) é opcional. Neste capítulo, são discutidas algumas técnicas de processamento e extração de características que foram utilizadas nos experimentos realizados para validar o trabalho desenvolvido nesta tese.

2.2 Segmentação de Imagens

A segmentação consiste na subdivisão de uma imagem em regiões distintas, levando em consideração as propriedades de descontinuidade e homogeneidade da imagem. Ela consiste em

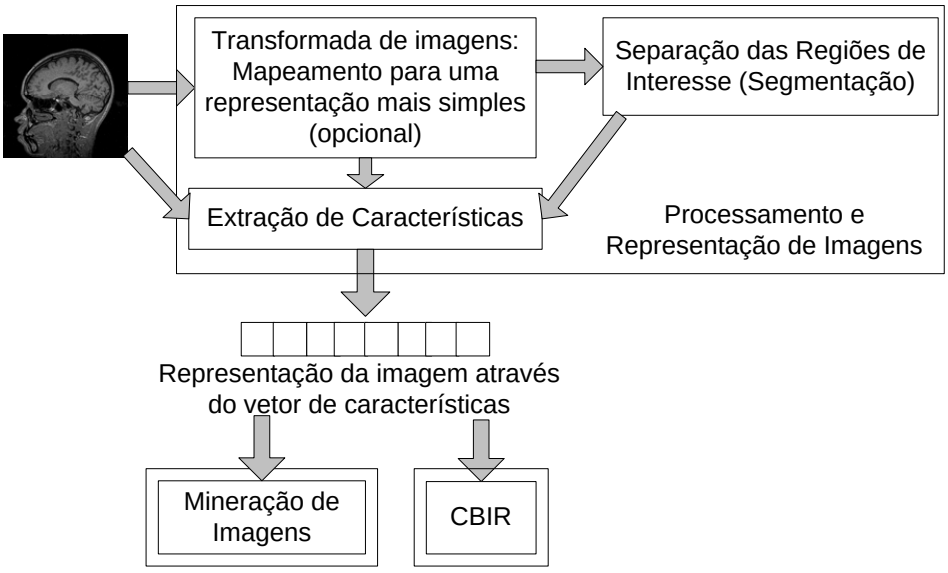


Figura 2.1: Processamento e representação de imagens como primeira etapa da mineração de imagens e do processamento nos sistemas CBIR.

uma das principais etapas para a análise automática de imagens, pois a partir dela podem ser delimitados os objetos sobre os quais se deseja extrair padrões.

Existem na literatura diversas taxonomias para classificação dos métodos de segmentação de imagens. Aqui é utilizada a classificação adotada em [Gonzalez & Woods, 2008] que divide os métodos de segmentação em três categorias básicas: limiarização (*thresholding*); segmentação baseada em bordas; e segmentação baseada em regiões. Recentemente outras classificações estão surgindo, como a classificação que leva em consideração cores [Al Aghbari & Al Haj, 2006].

Nesta seção são apresentados alguns dos principais métodos de segmentação de imagens encontrados na literatura, incluindo os métodos utilizados nos experimentos desta tese.

2.2.1 Limiarização (*thresholding*)

Um dos métodos mais simples de segmentação de imagens é a limiarização. Nesta técnica todos os pixels que estão dentro de uma mesma faixa de intensidade são classificados como pertencentes a uma mesma região. Em sua forma mais geral a limiarização pode ser descrita matematicamente pela equação:

$$S(i,j) = k \quad \text{se} \quad T_{k-1} \leq f(i,j) < T_k \quad \text{para} \quad k = 1,2,...,m$$

onde $S(i,j)$ é a função resultante, $f(i,j)$ é a função original (imagem), $T_0,...,T_m$ são os limiares *thresholds* e m é o número de classes distintas a serem aplicadas à imagem. No caso particular de $m = 2$, o método de limiarização é denominado *limiarização binária*.

Um dos problemas do método de limiarização é como determinar os valores dos limiares (T_i). Uma maneira de determinar esses valores é utilizar a seleção manual, onde o usuário com base na visualização do resultado da segmentação seleciona os valores dos limiares.

Grande parte das técnicas automáticas de limiarização se baseiam no histograma que, para imagens monocromáticas, é uma função da freqüência de ocorrência de um determinado nível de cinza dentro da imagem. Uma técnica bastante utilizada é a determinação dos limiares como pontos de mínimo do histograma da imagem (vales). No entanto, se os vales são longos, a escolha dos limiares passa a ser arbitrária. Além disso, esse método é bastante sensível a ruídos. Uma técnica mais eficiente é a técnica de Otsu [Otsu, 1979]. A técnica de Otsu se baseia na escolha do valor de corte que maximiza a medida de variância entre duas partes do histograma, ou seja, o objetivo é encontrar T que minimize a função:

$$u(T) = q_1(T)\sigma_1^2(T) + q_2(T)\sigma_2^2(T)$$

onde $q_1(T)$ é o número de pixels com intensidade menor que T; $q_2(T)$ é o número de pixels com intensidade maior que T; $\sigma_1^2(T)$ é a variância dos pixels com intensidade menor que T; e $\sigma_2^2(T)$ é a variância dos pixels com intensidade maior que T.

Um exemplo de resultado da aplicação da técnica de Otsu é apresentado na figura 2.2.

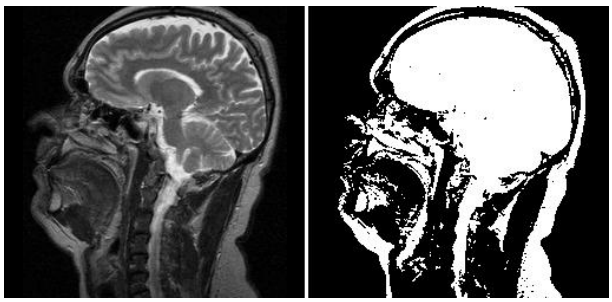


Figura 2.2: Exemplo de uma imagem e sua segmentação pelo método de Otsu.

2.2.2 Segmentação Baseada em Bordas

A segmentação baseada em bordas é usada para encontrar descontinuidades significativas nos níveis de cinza da imagem. Uma borda é um limite entre duas regiões com propriedades relativamente distintas de níveis de cinza. Após a detecção das bordas, geralmente é empregado um método que conecta os fragmentos da borda (método de *enlace*), gerando os contornos dos objetos.

A maioria das técnicas de detecção de bordas usa um operador local diferencial. Em imagens digitais geralmente uma borda é modelada como uma transição suave de níveis de cinza em vez de uma transição abrupta. A figura 2.3 (a) mostra uma imagem com uma faixa clara sobre um fundo escuro, o perfil de cinza ao longo de uma linha de varredura horizontal e a

primeira e a segunda derivada desse perfil. A figura 2.3 (b) tem as mesmas informações para uma imagem com uma faixa escura sobre um fundo claro. Observe na figura 2.3 (a) que o sinal da primeira derivada inverte sempre que uma borda é atingida e que o sinal da segunda derivada sempre é positivo do lado da borda mais escuro. Assim, através da primeira derivada é possível detectar as bordas de um objeto e através da segunda derivada é possível detectar se um pixel está do lado mais claro ou mais escuro da borda.

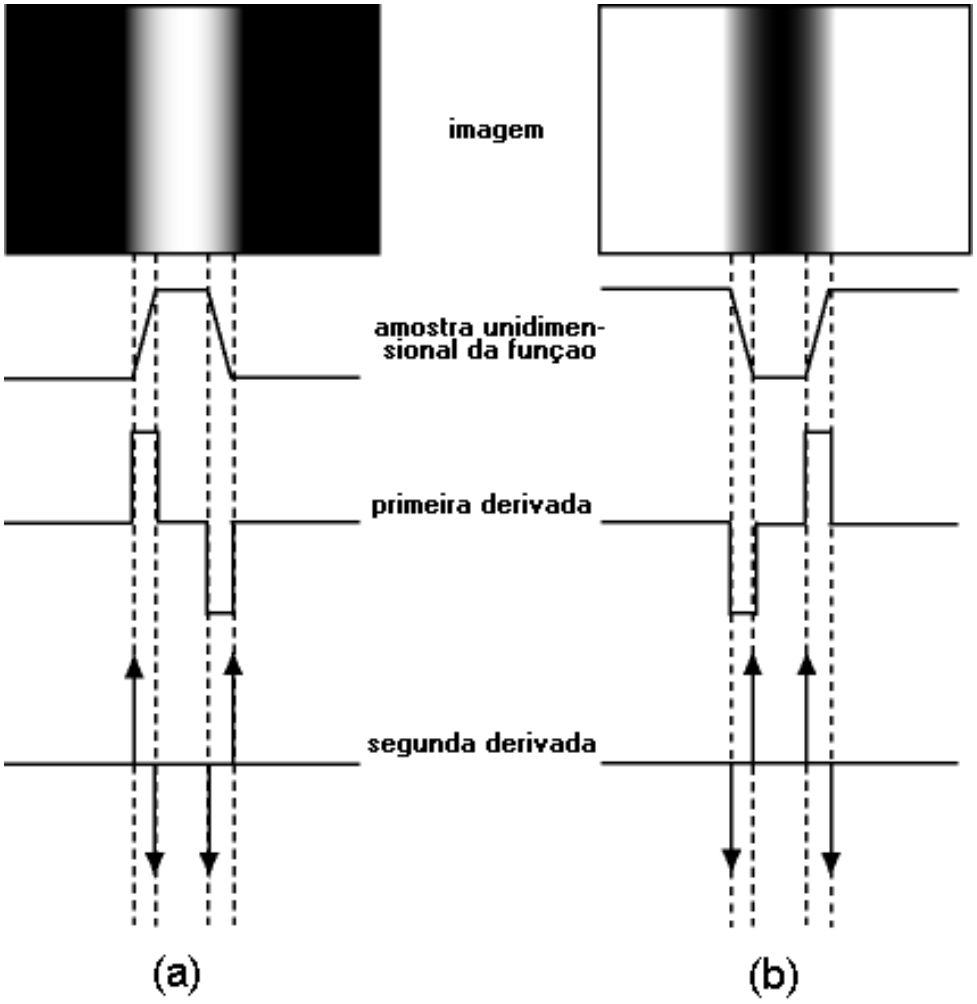


Figura 2.3: Detecção de bordas através de operadores de derivação: (a) faixa clara sobre fundo escuro; (b) faixa escura sobre o fundo claro.

O conceito de gradiente é utilizado para a diferenciação de imagens. O gradiente de uma imagem $f(x,y)$ na posição (x,y) é dado pelo vetor:

$$\nabla f(x,y) = G[f(x,y)] = \begin{bmatrix} G_x \\ \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

O vetor gradiente aponta para a direção de mudança mais rápida de f na posição (x,y). A

tarefa de detecção de bordas requer somente a magnitude do vetor gradiente, que geralmente é chamada simplesmente de gradiente:

$$|G[f(x,y)]| = \sqrt{G_x^2 + G_y^2}$$

A direção do vetor gradiente é usada para conectar contornos e é obtida pela equação:

$$\theta(x,y) = \arctan\left(\frac{G_y}{G_x}\right)$$

onde o ângulo θ é medido com relação ao eixo x.

A derivação pode ser implementada de diferentes formas. O uso dos operadores de Sobel para essa tarefa possui a vantagem de fornecer um efeito de suavização, o que é importante uma vez que a derivação aumenta o ruído. A figura 2.4 mostra as máscaras usadas para o cálculo do gradiente no ponto central de uma região 3×3 de uma imagem. A figura 2.4 (a) mostra o operador de Sobel usado para o cálculo de G_y e a figura 2.4 (b) mostra o operador de Sobel usado para o cálculo de G_x.

1	2	1
0	0	0
-1	-2	-1

(a)

-1	0	1
-2	0	2
-1	0	1

(b)

Figura 2.4: Operadores de Sobel (a) máscara usada para computar G_y; (b) máscara usada para computar G_x.

A figura 2.5 (a) mostra uma imagem e os itens (b) e (c) mostram respectivamente os resultados da aplicação do operador de Sobel para a obtenção de G_x e G_y. Observe que a aplicação de G_x realça contornos horizontais enquanto a aplicação de G_y realça contornos verticais.

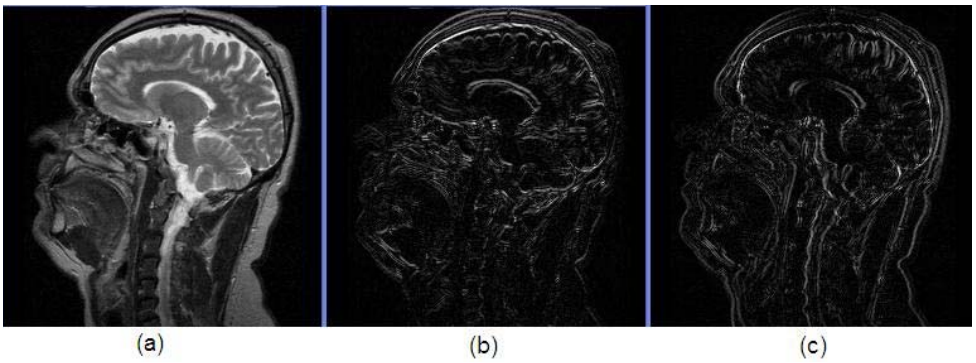


Figura 2.5: (a) Imagem original e os resultados da aplicação dos operadores de Sobel para a obtenção de (b) G_x e (c) G_y.

Para encontrar a derivada de segunda ordem de uma função $f(x,y)$ é usado o operador Laplaciano, que é definido por:

$$\Delta^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

Uma das maneiras de computar o operador Laplaciano é utilizar a máscara apresentada na figura 2.6.

0	-1	0
-1	4	-1
0	-1	0

Figura 2.6: Máscara usada para computar o operador Laplaciano.

Por ser uma derivada de segunda ordem, o Laplaciano é bastante sensível a ruídos, sendo usado na função de determinar se um pixel está do lado claro ou escuro da imagem e para a localização de bordas usando a propriedade de cruzamentos por zero detalhada em [Gonzalez & Woods, 2008].

Um método mais preciso para detectar bordas é o modelo de contorno ativos (*snakes*) detalhado em [Kass et al., 1987]. As *snakes* são contornos que são atraídos para as bordas dos objetos da imagem e adaptam-se a essas bordas em um processo de convergência.

Os algoritmos de detecção de bordas devem ser seguidos por procedimentos de ligação para juntar e organizar os pixels da borda em fronteiras significativas. Os procedimentos mais conhecidos de ligação são os métodos baseados no processamento local, na transformada de *Hough* e em grafos. O processamento local é a abordagem mais simples que consiste na análise das características dos pixels dentro de uma pequena vizinhança (3×3 ou 4×4, por exemplo) de cada ponto de uma imagem previamente submetida a um processo de detecção de bordas. Os pontos que possuem características similares são interligados. As características analisadas são intensidade do gradiente $|G[f(x,y)]|$ e a direção do gradiente da imagem ($\theta(x,y)$). Dois pixels vizinhos (*x,y*) e (*x',y'*) são similares se $|\Delta f(x,y) - \Delta f(x',y')| \leq T$ e se $|\theta(x,y) - \theta(x',y')| \leq A$, onde T e A representam respectivamente um limiar e um ângulo limite pré-estabelecidos.

O método baseado na transformada de *Hough* para ligação de bordas considera relações globais entre os pixels e é utilizado para encontrar curvas específicas dentro da imagem, como circunferências e retas. Isso é feito mapeando a imagem do espaço original cartesiano para o espaço dos parâmetros da curva procurada.

Os métodos baseados em grafos para ligação de bordas representam segmentos de borda na forma de um grafo e buscam, através desse grafo, os caminhos de menor custo que correspondem às bordas. Esse método fornece uma representação robusta que funciona bem na presença de ruídos. No entanto, o problema de achar o caminho de custo mínimo requer um alto poder computacional. Por isso, uma abordagem comum é não usar uma solução ótima, mas sim utilizar heurísticas para reduzir o esforço computacional da busca pela solução [Costa & Jr., 2001].

2.2.3 Segmentação Baseada em Regiões

A segmentação baseada em regiões tem como objetivo particionar a imagem em regiões significativas. As principais técnicas usadas nessa abordagem são a técnica de crescimento de regiões, a técnica de divisão e fusão de regiões e as técnicas de alagamento.

O crescimento de regiões é uma técnica que agrupa sub-regiões em regiões maiores. A abordagem mais simples para a técnica de crescimento de regiões é a agregação de pixels. Nessa técnica, alguns pontos iniciais são escolhidos (sementes) para representar as regiões. Pixels vão sendo anexados às regiões de acordo com a similaridade que eles possuem com as regiões. Um critério simples para a agregação de um pixel à uma região é o módulo da diferença dos tons de cinza da semente e do pixel. Nesta técnica, existem dificuldades para a escolha adequada das sementes e para a escolha adequada do critério de inclusão dos pixels nas regiões. O conhecimento prévio de algumas características das regiões, como o formato e tamanho médio das regiões, pode ser usado para tornar a condição de parada do algoritmo mais precisa e agilizar o processo de segmentação.

Na técnica de divisão e fusão de regiões, a imagem é recursivamente particionada e agrupada de acordo com uma condição pré-estabelecida. Por exemplo, considere R_i uma região de uma imagem e P um predicado. R_i é subdivido em quatro novas regiões (quadrantes) se $P(R_i)$ for falso. Duas regiões adjacentes R_i e R_j são fundidas quando $P(R_i \cup R_j)$ for verdadeiro. Essa técnica possui uma representação conveniente na estrutura *quadtree*, que é uma árvore onde um nó não folha sempre possui quatro descendentes. A figura 2.7 mostra um exemplo de particionamento de uma imagem e a *quadtree* associada a ela.

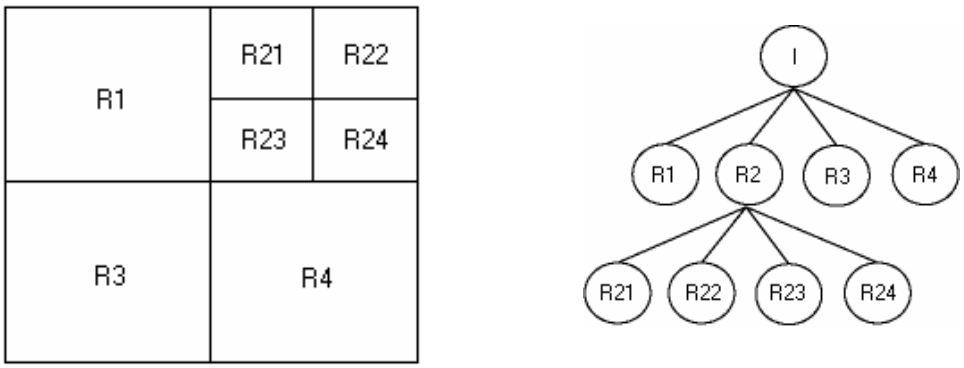


Figura 2.7: Exemplo de representação de uma imagem: a imagem particionada (à esquerda) e sua representação em *quadtree* (à direita).

A figura 2.8 ilustra um exemplo do funcionamento do algoritmo de divisão e fusão, onde o predicado $P(R_i)$ é verdadeiro se todos os pixels da região possuírem a mesma intensidade e falso, caso contrário. Inicialmente, a imagem é dividida em 4 regiões (figura 2.8 (a)). Posteriormente, cada nova região da imagem é dividida em 4 regiões (figura 2.8 (b)) e, finalmente,

regiões adjacentes de mesma intensidade são fundidas (figura 2.8 (c)).

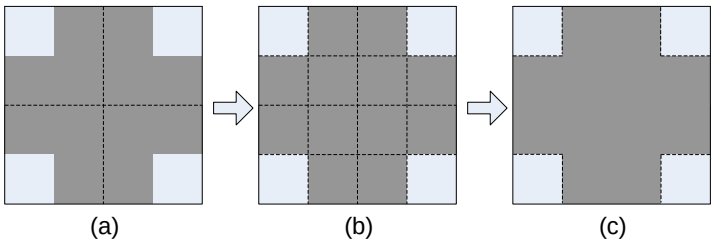


Figura 2.8: Exemplo de funcionamento da técnica de divisão e fusão: (a) primeira divisão em 4 regiões; (b) subdivisão das regiões; (c) fusão das regiões adjacentes.

O conceito de segmentação por textura também pode ser aplicado na segmentação por região, onde o predicado usado para decidir se uma região será ou não subdividida baseia-se em características de textura da região, como a média e o desvio padrão das intensidades dos pixels.

Existe também a transformada de *watersheds* [Vincent & Soile, 1991], uma técnica de segmentação pertencente ao campo da morfologia matemática, que também pode ser considerada um tipo de segmentação por região. *Watersheds* são linhas divisoras de água (LDA). A idéia básica dessa transformada é considerar que uma imagem em níveis de cinza possa ser representada por uma superfície topográfica em que os níveis indicam a altitude do ponto no relevo. Quando uma gota de água cai nessa superfície, ela converge para uma região de mínimo chamada de represa. No entanto, em alguns pontos não é possível determinar para onde irá escorrer a gota de água que ali cair. Esses pontos são as linhas divisoras de água, que são as fronteiras das represas retornadas pelo método. Freqüentemente, a transformada de *watershed* é aplicada sobre o gradiente de uma imagem, para que os contornos da imagem sejam realçados antes do processo de segmentação.

2.2.4 Segmentação utilizando o Método EM/MPM

Freqüentemente os objetos presentes em imagens do mundo real são caracterizados por microtexturas de comportamento não determinístico (aleatório), por isso, abordagens probabilísticas de segmentação têm sido bastante difundidas. De uma maneira geral, os algoritmos estatísticos de segmentação definem rótulos nos pixels de uma imagem como variáveis aleatórias que formam um campo aleatório bidimensional. Os **Campos Aleatórios de Markov** (*Markov Random Fields*) são bastante adequados para a modelagem de microtexturas, pois definem uma função de probabilidade por meio de características locais de vizinhança [Gerhardinger, 2006]. A representação de cada objeto da imagem pode ser feita construindo um campo de rótulos (classes). No modelo estocástico, a imagem é o dado observado e o campo de rótulos é o dado ausente.

O EM/MPM (***E**xpectation **M**aximization / **M**aximization of the **P**osterior **M**arginals*) é um método não-supervisionado de segmentação de imagens. Nesta técnica são utilizados dois mo-

delos: um Modelo de Mistura de Gaussianas (GMM) referente a imagem observada (em tons de cinza); e um campo aleatório de Markov (*Markov Random Field* - MRF) para o mapa de classificação dos pixels, que é o resultado da segmentação. O objetivo do algoritmo é estimar os parâmetros do modelo GMM, por meio do método EM, e minimizar o número estimado de pixels classificados erroneamente no mapa de classes por meio da classificação Bayesiana. A abordagem Bayesiana para este caso consiste, inicialmente, em eleger uma função densidade de probabilidade para o mapa de classes até então desconhecido. Essa função é denominada probabilidade prévia (*prior probability*), pois ela não é baseada em dados existentes. A partir dessa distribuição deseja-se obter uma nova função, denominada probabilidade posterior (*posterior probability*). A probabilidade posterior, neste caso, é uma função de custo que indica o quão correto está o mapa de classes atual em relação ao resultado esperado. O tão conhecido teorema de Bayes postula que a probabilidade posterior é proporcional à probabilidade prévia vezes uma outra função denominada **função de verossimilhança** (*likelihood function*).

O método proposto por Balan [Balan, 2007] otimiza o método EM/MPM [Comer & Delp, 2000], onde o parâmetro de interação espacial do modelo de Markov é ajustado para uma convergência mais rápida. A variação está na aplicação da técnica de *annealing* ao algoritmo de segmentação EM/MPM, que consiste em aumentar gradativamente o valor do parâmetro de interação espacial do modelo de Markov (β) durante a segmentação de uma imagem. A imagem é segmentada de acordo com um mapa de classificação dos pixels, onde cada pixel é classificado com um valor de classe $k \in [1, L]$. Os pixels classificados com a mesma classe c formam grupos onde todos os pixels estão direta ou indiretamente conectados uns aos outros por um determinado critério de vizinhança. A figura 2.9 apresenta um exemplo do resultado da segmentação obtida de uma imagem em cinco classes (contando o *background*), usando o método proposto por Balan. As regiões de textura obtidas também são mostradas separadamente para uma melhor visualização.

Ao extrair características das regiões segmentadas pelo método proposto por Balan é importante levar em consideração que os campos aleatórios de Markov expressam apenas propriedades locais das imagens. Assim, é importante extrair propriedades globais das imagens para discriminá-las, como as propriedades fractais (descritas no capítulo 4) das regiões. Além disso, ao se trabalhar com imagens de ressonância magnética, como nos estudos de caso desta tese, pode-se considerar que a padronização da configuração do equipamento e posicionamento do paciente seja razoável, o que permite que características variantes as transformações geométricas de escala, translação e rotação sejam também utilizadas para a representação das regiões segmentadas.

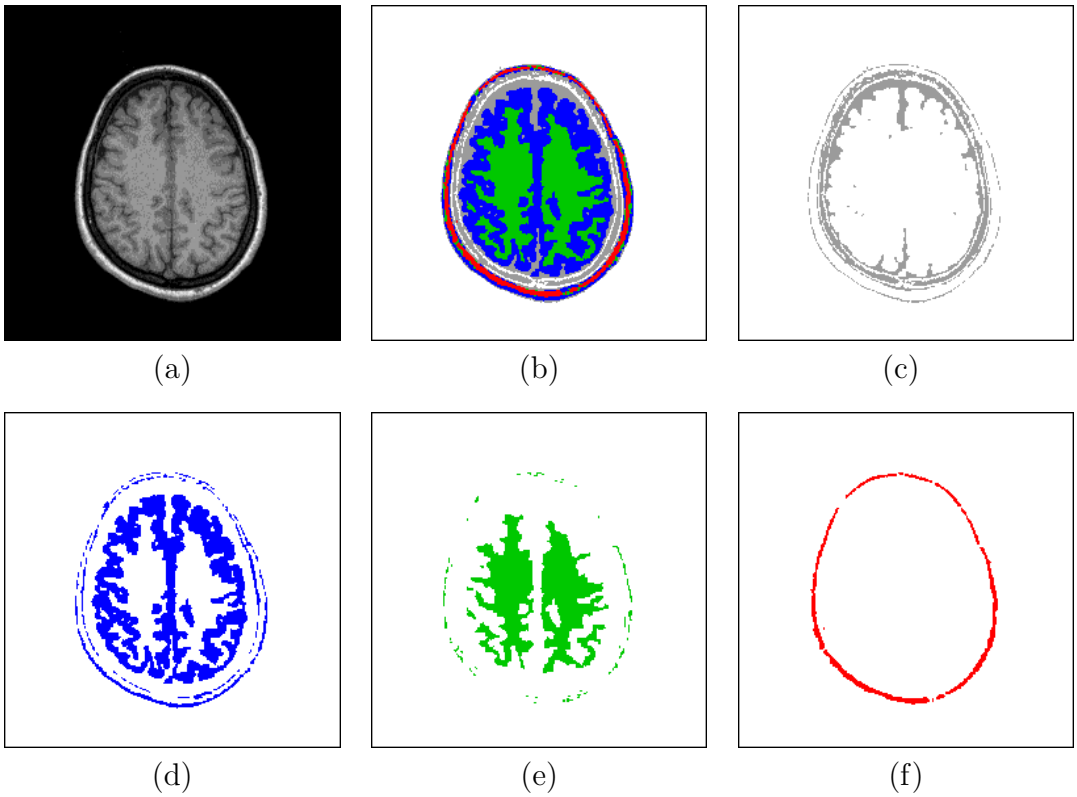


Figura 2.9: Exemplo de segmentação realizada pelo método proposto por Balan. (a) Imagem original; (b) Imagem segmentada em 5 classes (incluindo o *background*); (c) região da classe 1 (fluido cérebro-espinhal); (d) região da classe 2 (massa cinzenta); (e) região da classe 3 (massa branca); (f) região da classe 4 (dura, medula óssea, gordura) [Balan et al., 2007].

2.3 Características para Representar as Imagens

Para a obtenção da representação, as imagens podem ser submetidas a um pré-processamento (transformada de imagens e/ou segmentação) ou diretamente submetidas a extratores de características, que são algoritmos que extraem informações intrínsecas da imagem (textura, cor e forma). Essas características são utilizadas para formar a representação das imagens através de vetores de características. Esses vetores de características são utilizados no lugar das imagens no processo de indexação, recuperação e mineração de imagens. Nesta seção, são descritos alguns dos principais extratores utilizados para a extração de características de imagens, incluindo aqueles utilizados nos experimentos desta tese.

2.3.1 Assinaturas de forma

As assinaturas de forma são funções unidimensionais discretas que representam o contorno de um objeto de uma imagem. Idealmente uma assinatura de forma descreve de maneira unívoca a forma do objeto ao qual ela está associada. No entanto, antes de obter a assinatura de forma

é necessário que a borda ou o contorno do objeto a ser descrito tenha sido determinado em uma etapa prévia de segmentação, gerando um vetor de coordenadas discretas $(x(t), y(t))$, onde $t = \{0, 1, ..., N-1\}$, e N representa o número total de amostras deste contorno. Uma assinatura de forma bastante utilizada é a função de distância ao centróide da região.

A função distância do centróide $r(t)$ é expressa pela distância entre os pontos que correspondem às amostras do contorno do objeto $(x(t), y(t))$ e o centróide (ou centro de massa) do objeto (x_c, y_c):

$$r(t) = \sqrt{(x(t) - x_c)^2 + (y(t) - y_c)^2}, \quad \text{onde,} \quad x_c = \frac{1}{N}\sum_{t=0}^{N-1} x(t), \quad y_c = \frac{1}{N}\sum_{t=0}^{N-1} y(t)$$

A função $r(t)$ é inerentemente invariante à translação do objeto. Uma operação de rotação do objeto causa em $r(t)$ um deslocamento circular, e uma transformação de escala altera $r(t)$ linearmente.

2.3.2 Histograma

As cores presentes em uma imagem possuem um papel bastante significativo na indexação e recuperação da mesma. O histograma, que apresenta o número de pixels de uma imagem para cada cor (ou nível de cinza em imagens monocromáticas), é freqüentemente usado para representar a imagem através da sua distribuição de cor. O histograma é invariante às operações de translação e rotação. Muitas vezes, para aumentar a eficiência do processamento, as cores da imagem são re-quantizadas, para diminuir o número de cores possíveis e facilitar o tratamento das mesmas através de seu histograma. A re-quantização do histograma comprime um intervalo de valores de intensidade em um único valor ("quantum"). A figura 2.10 apresenta uma imagem com 256 níveis de cinza e seu histograma.

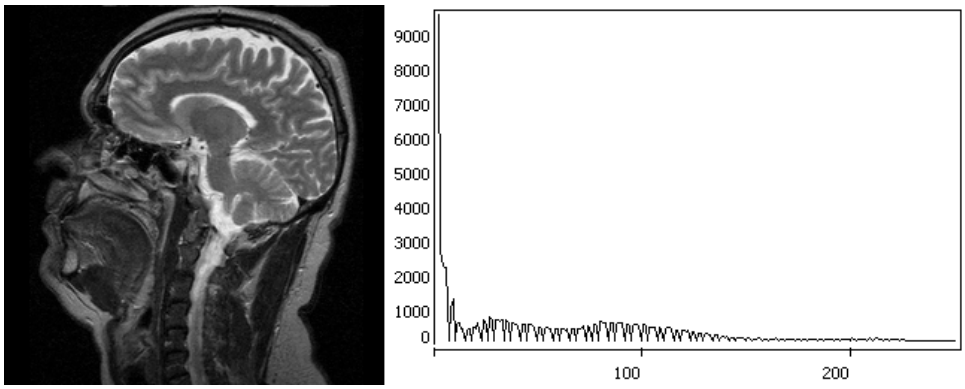


Figura 2.10: Exemplo de imagem e seu histograma.

A normalização do histograma geralmente é feita considerando que o número total de pixels da imagem equivale ao valor 1. O valor de cada *bin* é a fração de pixel que ocorreu nessa faixa de intensidade. O histograma normalizado é invariante a transformações geométricas, bem

como a transformações lineares de luminosidade. Outra operação que pode ser feita sobre o histograma é a equalização. A equalização visa obter a máxima variância do histograma da imagem alcançando assim um maior contraste.

Dois histogramas de cores podem ser comparados pelo somatório das diferenças absolutas ou quadráticas sobre o número de pixels de cada cor. Tal esquema é bastante simples e tolerante a pequenas alterações na imagem. Histogramas baseados em cores têm sido bastante usados em sistemas de recuperação de imagens por conteúdo, tanto acadêmicos [Ko et al., 2000] [Pass et al., 1996] quanto comerciais, como QBIC [Flickner & alli, 1995].

A popularidade da utilização de histogramas de cores em sistemas de recuperação de imagens por conteúdo deve-se principalmente a três fatores [Pass et al., 1996]: (a) ser computacionalmente simples e barato de calcular; (b) pequenas alterações de movimentação na imagem pouco afetam o histograma; e (c) objetos distintos freqüentemente possuem histogramas diferentes. No entanto, não é possível separar ou reconhecer imagens utilizando apenas o histograma das mesmas, pois duas ou mais imagens bastante diferentes podem ter histogramas semelhantes. A Figura 2.11 apresenta 4 exemplos de imagens que possuem o mesmo histograma.

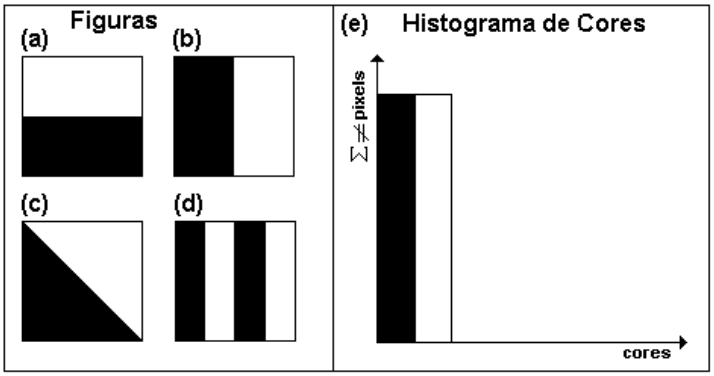


Figura 2.11: Exemplo de imagens ((a),(b),(c) e (d)) com o mesmo histograma (e).

Além do problema da ambigüidade, o histograma também apresenta o problema de ter alta dimensionalidade. Como o número de cores é grande (geralmente mais de 100 níveis), indexar vetores com essa dimensão é algo problemático. Isso porque um histograma para 100 cores distintas pode ser visto como um ponto no espaço 100-dimensional e, para valores dessa ordem, a maior parte das estruturas de índices espaciais sofre com a "maldição da alta dimensionalidade" [Pagel et al., 2000], onde o melhor método de acesso passa a ser a busca seqüencial.

Para contornar o problema da maldição da alta dimensionalidade, em [Bueno, 2002] foi proposto o *histograma métrico*, onde somente os valores de máximo e mínimo do histograma são armazenados para compor o vetor de características, de dimensionalidade variável, para representar as imagens. Em um **histograma métrico**, o equivalente ao *bin* do histograma tradicional

é denominado *bucket*, onde estes correspondem a um segmento de reta na aproximação do histograma normalizado, ou seja, os *buckets* na verdade correspondem a um subconjunto de *bins* do histograma original. Portanto, dessa maneira, é possível limitar o formato do histograma e também reduzir a dimensionalidade deste. Na Figura 2.12 são ilustrados os *bins* e os *buckets* de um dado histograma. O *histograma métrico* apresenta um ganho efetivo de desempenho, considerando o processo de indexação e recuperação de imagens. No entanto, esse tipo de histograma tem a dimensionalidade variável, sendo aplicado somente em estruturas de indexação métricas.

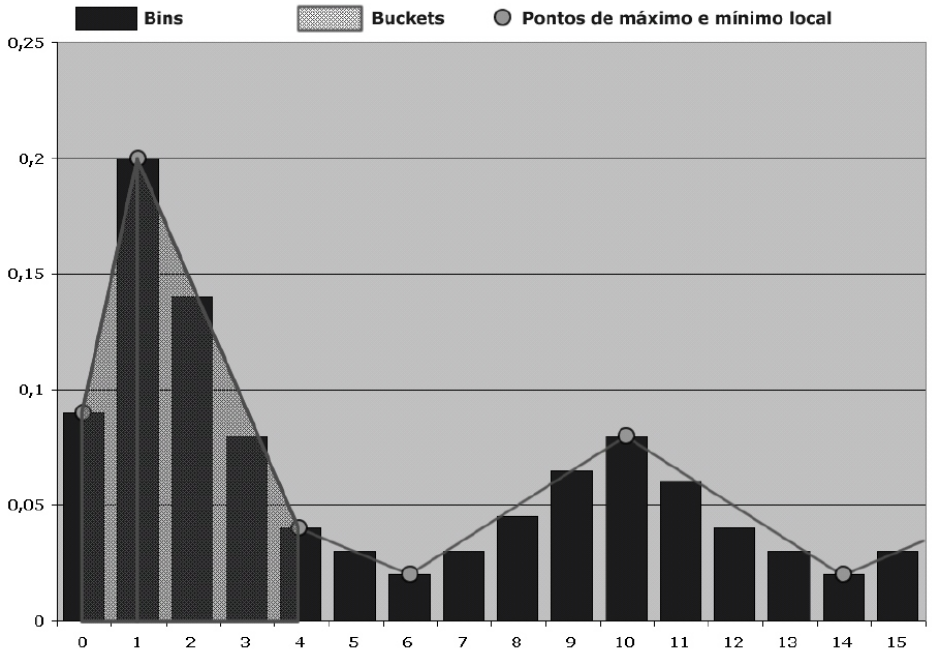


Figura 2.12: Histograma normalizado com pontos de controle de máximo e mínimo local, os quais definem os *buckets* correspondentes ao seu *histograma métrico* [Bueno, 2002].

2.3.3 Momentos Invariantes

Os momentos invariantes promovem uma caracterização global de forma de uma imagem. O momento *2D* de ordem $(p+q)$ de uma imagem digital $f(x,y)$ é definido como:

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y)$$

para *p,q* ={0, 1, 2,...}, onde os somatórios são executados percorrendo todas as coordenadas x e y da imagem. O momento de ordem 0 (m_{00}) representa a superfície, enquanto os momentos de ordem 1 (m_{01}) e (m_{10}) definem o centro da gravidade $\overline{x}$ e $\overline{y}$ da imagem. Assim, o *momento central* é definido como:

$$\mu_{pq} = \sum_y \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y), \text{ onde } \bar{x} = \frac{m_{10}}{m_{00}} \text{ e } \bar{y} = \frac{m_{01}}{m_{00}}$$

Os momentos centrais são invariantes a rotação e translação. Para transformá-los em invariantes a escala, os mesmos devem ser normalizados pelo tamanho da imagem. Os momentos centrais normalizados de ordem $(p+q)$, também chamados de *momentos de Hu*, são definidos como:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}, \text{ onde } \gamma = \frac{p+q}{2} + 1 \text{ para } p,q = \{0,1,2,...\} \text{ e } p+q = \{2,3,...\}$$

Os sete *momentos de Hu* são invariantes a translação, escala e rotação e podem ser derivados das equações acima:

$\phi_1 = \eta_{20} + \eta_{02}$

$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$

$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$

$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2$

$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})^2$

$\phi_7 = 3(\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} - \eta_{03})]$

2.3.4 Momentos de Zernike

Os momentos de Zernike são utilizados para representar características de forma das imagens. Os polinômios de Zernike, dizem respeito a um conjunto de polinômios complexos que formam uma base ortogonal no interior de um círculo de raio unitário $x^2 + y^2 \leq 1$ [Khotanzad & Hong, 1990]. Esses polinômios podem ser formalmente denotados por meio das coordenadas polares através da equação 2.1, onde n é um valor inteiro, maior ou igual a zero, que define a ordem do polinômio; e m, trata-se também de um valor, que pode assumir valores tanto positivos quanto negativos, o qual descreve a dependência angular, ou rotação, do polinômio.

$$V_{n,m}(x,y) = V_{n,m}(\rho\ cos(\theta), \rho\ sin(\theta)) = R_{n,m}(\rho)e^{im\theta} \tag{2.1}$$

Considerando os valores n e m, estes devem satisfazer às seguintes condições:

$$n - |m| \quad \text{deve ser par; e} \quad -n \leq m \leq n \tag{2.2}$$

Para realizar o cálculo dos momentos de Zernike, a imagem (ou região de interesse) deve primeiramente ser mapeada das coordenadas do plano cartesiano para o disco unitário por meio

das coordenadas polares, onde o centro da imagem é a origem do disco unitário. Os pixels da imagem mapeados para fora do disco não são utilizados no cálculo dos momentos. Para realizar tal mapeamento das coordenadas cartesianas (x,y) para o sistema polar, calcula-se: $\rho = \sqrt{x^2+y^2}$ e $\theta = arctan(\frac{y}{x})$, onde ρ diz respeito à distância do ponto (x,y) à origem, e θ denota o ângulo formado entre o vetor da norma (comprimento) ρ e o eixo x, no sentido anti-horário. A Figura 2.13 ilustra tal representação, onde a origem é denotada por C, o par (x,y) são as coordenadas cartesianas do ponto P e (ρ,θ) suas coordenadas polares, sendo que $x = \rho * cos\theta$ e $y = \rho * sen\theta$.

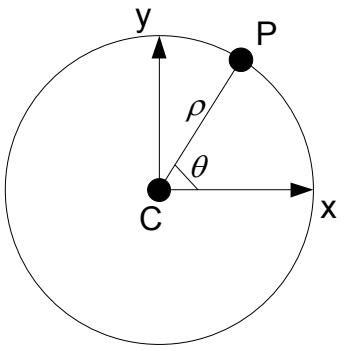


Figura 2.13: Representação das coordenadas polares de um ponto P sobre o plano euclidiano cuja origem é denotada por C.

A função $R_{n,m}(\rho)$ é denominada polinômio radial e pode ser definida formalmente como:

$$R_{n,m}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s \frac{(n-s)!}{s!\,(\frac{n+|m|}{2}-s)!\,(\frac{n-|m|}{2}-s)!}\,\rho^{n-2s} \tag{2.3}$$

Pode-se notar que $R_{n,m}(\rho) = R_{n,-m}(\rho)$. Caso as condições descritas pelas restrições da equação 2.2 não sejam satisfeitas, o polinômio $R_{n,m}(\rho)$ é nulo. Os primeiros seis polinômios radiais são:

$$\begin{aligned} R_{0,0}(\rho) &= 1 & R_{1,1}(\rho) &= \rho \\ R_{2,0}(\rho) &= 2\rho^2 - 1 & R_{2,2}(\rho) &= \rho^2 \\ R_{3,1}(\rho) &= 3\rho^3 - 3\rho & R_{3,3}(\rho) &= \rho^3 \end{aligned}$$

Os momentos de Zernike representam uma projeção de uma dada imagem $f(x,y)$ sobre a base ortogonal formada pelos polinômios $V_{n,m}(x,y)$. O momento de Zernike $Z_{n,m}$ de ordem n e repetição (ou dependência angular) m de uma imagem digital é definido pela equação 2.4, onde $V^*_{n,m}$ denota o conjugado do valor complexo de $V_{n,m}$.

$$Z_{n,m} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y)\, V^*_{n,m}(x,y), \qquad x^2 + y^2 \leq 1 \tag{2.4}$$

Segundo as condições da equação 2.2, para cada ordem n existem $\frac{n}{2}$ momentos com m dependência angulares (repetição) distintas. Dessa forma, existem $(n+1)(n-1)/2$ momentos de ordem inferior ou igual a n. Tem-se que a relação entre os momentos de ordem n é $Z^*_{n,m} = Z_{n,-m}$.

Os momentos de Zernike são invariantes a transformação de rotação na imagem. Para que os momentos de Zernike também sejam invariantes à operação de translação é necessário que a origem do disco unitário coincida com o centro de massa da imagem no mapeamento. Já a invariância à escala é obtida escalonando-se a imagem de modo que a sua massa ($m_{0,0}$) passe a assumir um valor β definido anteriormente. A invariância à escala e a invariância à translação são obtidas através da equação 2.5, onde (x_c, y_c) denota o centro de massa do objeto.

$$h(x,y) = f\left(\frac{x}{\alpha} + x_c,\ \frac{y}{\alpha} + y_c\right) \quad \text{onde} \quad \alpha = \sqrt{\frac{\beta}{m_{0,0}}} \tag{2.5}$$

O centro de massa é (x_c, y_c) é calculado a partir da formulação $m_{p,q} = \sum_x \sum_y x^p y^q$ tal que:

$$x_c = \frac{m_{1,0}}{m_{0,0}} \quad \text{e} \quad y_c = \frac{m_{0,1}}{m_{0,0}}$$

Nesta tese utilizou-se os momentos de Zernike para representar imagens de mamografia. Os tumores de mama são fortemente relacionados com a forma das lesões. Os momentos de Zernike são apontados na literatura como um descritor bastante adequado para descrever características de forma. Uma vantagem do seu uso é a não necessidade de uma etapa prévia de segmentação das imagens. De fato, a maioria dos trabalhos da literatura trabalham com imagens previamente submetidas a um processo de segmentação manual [Rangayyan et al., 2000] ou trabalham com imagens sem segmentação [Felipe et al., 2006], isso porque a segmentação automática das lesões de mama é uma tarefa bastante difícil devido a semelhança existente entre as lesões e o tecido sadio da mama nas imagens.

2.3.5 Textura

O que é textura? Como definir textura? "Textura se refere à repetição de elementos básicos da imagem chamados *textels*. A distribuição dos *textels* pode ser periódica ou aleatória. Texturas naturais geralmente possuem um comportamento aleatório, sendo que as artificiais possuem um comportamento periódico e determinístico" [Jain, 1993]. As medidas de textura capturam essencialmente a granularidade e padrões repetitivos na distribuição dos pixels. Por exemplo, vidro, tijolos, grama, madeira e papel diferem entre si tanto pela suavidade da textura quanto pela repetição de padrões [Balan, 2007].

Algumas das mais conhecidas técnicas de extração de características de textura são as *Wavelets*, os filtros de Gabor e aquelas baseadas nas "matrizes de co-ocorrência". A partir dessas técnicas é possível computar medidas de periodicidade, granularidade, direcionalidade e regularidade das regiões das imagens. Vários experimentos realizados nesta tese foram obtidos aplicando os descritores de Haralick sobre as matrizes de co-ocorrência.

As matrizes de co-ocorrência, também denominadas matrizes SGLD (*Spatial Gray Level Dependence*) [Haralick et al., 1973], são uma das mais populares fontes de características de textura para imagens. Dado uma imagem f com um conjunto discreto de tons de cinza I, define-se a matriz de co-ocorrência $P_{d,\phi}(i,j)$, onde cada elemento (i,j) é um número inteiro que indica quantas vezes um pixel p_1 de nível de cinza i aparece distante de um pixel p_2 de intensidade j por uma distância d e um ângulo ϕ. As figuras 2.14 (b) e (c) ilustram duas matrizes de co-ocorrência para a imagem em tons de cinza representada na figura 2.14 (a). As matrizes de co-ocorrência são matrizes quadradas e simétricas em relação à diagonal principal, ou seja, $P_{d,\phi}(i,j) = P_{d,\phi}(j,i)$.

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

(a)

$$P_{1,0^\circ} = \begin{bmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \quad P_{1,135^\circ} = \begin{bmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{bmatrix}$$

(b) (c)

Figura 2.14: Exemplos de matrizes de co-ocorrência. (a) imagem; (b) matriz de co-ocorrência para o ângulo 0° e $d = 1$; (c) matriz de co-ocorrência para o ângulo 135° e $d = 1$.

Várias medidas podem ser extraídas de uma matriz de co-ocorrência. Haralick [Haralick et al., 1973] propõe 14, porém as mais utilizadas na literatura são as apresentadas na tabela 2.1.

Tabela 2.1: Características de textura de Haralick.

Característica	**Equação**	**Significado**
Variância	$\sum_i \sum_j (i-j)^2 P(i,j)$	Contraste
Entropia	$\sum_i \sum_j P(i,j) log(P(i,j))$	Suavidade
Energia	$\sum_i \sum_j P(i-j)^2$	Uniformidade
Homogeneidade	$\sum_i \sum_j \frac{P(i-j)}{(1+\|i-j\|)}$	Homogeneidade
3° Momento	$\sum_i \sum_j (i-j)^3 P(i,j)$	Distorção
Inverso da Variância	$\sum_i \sum_j \frac{P(i,j)}{(i-j)^2}$	Inverso do Contraste.

A medida de variância (contraste) analisa os valores da matriz com ênfase nos elementos mais distantes da diagonal, ou seja, os pontos cujos níveis de cinza possuem maior distinção entre si, o que corresponde a um indicador do nível de contraste da textura. As medidas de entropia e energia dão uma indicação do comportamento da textura em relação à sua periodicidade e uniformidade. A homogeneidade possui a mesma tendência, porém, com o sentido inverso. O 3° momento indica a distorção da imagem, enquanto o inverso da variância fornece uma noção de contraste.

A dimensão da matriz $P_{d,\phi}$ é definida pelo número de tons de cinza distintos na imagem. Freqüentemente algoritmos de quantização são aplicados na imagem antes do cálculo das matrizes de co-ocorrência, para a redução da quantidade de tons de cinza, proporcionando assim, a obtenção de matrizes de dimensões reduzidas.

Para se obter o conjunto de medidas que apresentam o maior poder de caracterização das imagens, é necessário computar um número razoável de matrizes para diferentes ângulos ϕ e distâncias d. Caso existam informações *a priori* das texturas a serem caracterizadas é possível tirar proveito disto. Por exemplo, para texturas mais refinadas é melhor que sejam computadas matrizes com parâmetros d pequenos, geralmente usando os valores 1 ou 2. Por outro lado, para texturas mais grosseiras é aconselhável a utilização de valores mais altos para d. Além disso, procura-se a utilização de ângulos múltiplos de $\pi/4$, já que a natureza espacial da imagem induz ao cálculo mais simplificado nas direções vertical, horizontal e nas diagonais. Vale notar que a seguinte relação é válida: $P_{d,\phi} = P_{d,\phi+\pi}$.

Como desvantagem da utilização das matrizes de co-ocorrência na extração de características de texturas pode-se citar seu alto custo de obtenção. Além disso, a escolha adequada dos parâmetros d e ϕ depende muitas vezes de um conhecimento prévio acerca do comportamento das imagens. No entanto, as características extraídas das matrizes de co-ocorrência são bastante utilizadas por promoverem uma descrição satisfatória das imagens sem a necessidade de uma etapa prévia de segmentação, que muitas vezes é extremamente cara em termos computacionais.

Neste trabalho, além dos descritores de Haralick descritos nesta seção, também é utilizado o seguinte descritor:

$$\textit{Step}\colon \sum_i \sum_j P(i,j)$$

Esse descritor é obtido em um passo intermediário para obter os demais descritores de Haralick. Ele fornece o valor acumulado das intensidades na imagem.

2.4 Considerações Finais

É importante o conhecimento do processamento das imagens para entender como é realizado o processo de recuperação de imagens por conteúdo, a mineração de imagens e também os

processos de análise automática de imagens que são discutidos nos próximos capítulos desta tese. Este capítulo apresentou uma breve descrição das principais técnicas de processamento de imagens que são empregadas para a obtenção da representação das imagens a partir de vetores de características. A maioria das técnicas apresentadas aqui são utilizadas para extrair características das imagens nos experimentos desta tese.

Capítulo 3

Sistemas de Apoio à Análise de Imagens Médicas

3.1 Considerações Iniciais

A tecnologia computacional tem evoluído muito nos últimos anos e isso permitiu um maior suporte à área médica. Ferramentas que permitem coletar e auxiliar análises de dados médicos têm sido desenvolvidas. Um desses suportes à área médica são os Sistemas de Informação Hospitalares (SIH) que armazenam informações a respeito do paciente, seus sintomas, diagnósticos e procedimentos médicos adotados. Em se tratando de imagens médicas, foram desenvolvidos os sistemas PACS (*Picture Archiving and Communication Systems*) [Morioka et al., 2005] onde dados de imagens médicas provenientes de exames como Raio-X, tomografia computadorizada, ressonância magnética e seus laudos são armazenados de forma integrada.

Outro suporte à área médica são os sistemas CAD (*Computer-Aided Diagnonis*) , que têm se tornado um dos principais tópicos de pesquisas da área de imagens médicas. Os sistemas CAD fornecem uma segunda opinião para os médicos radiologistas e estudantes de medicina, melhorando a precisão e a consistência dos diagnósticos radiológicos e reduzindo o tempo de leitura e análise das imagens. Por sua vasta aplicabilidade, é esperado que os sistemas CAD tenham um grande impacto na área de imagens médicas e exames radiológicos no século XXI [Doi, 2005]. O desenvolvimento de sistemas CAD envolve uma série de conceitos que vão desde representação e extração de características de imagens à mineração e descoberta de padrões.

Além dos sistemas PACS e CAD, os sistemas CBIR, que realizam busca por conteúdo em bases de imagens médicas, têm se mostrado bastante relevantes para a área médica. Nesses sistemas, radiologistas buscam imagens semelhantes a uma dada imagem de busca para obter dados históricos de imagens e exames semelhantes ao problema em questão, que os auxiliem no momento de fornecer o diagnóstico.

A busca de imagens tem sido realizada por meio de duas abordagens. A primeira é a associação de texto descritivo às imagens armazenadas no banco de dados de forma que a recuperação das imagens se baseie nessa descrição. Embora essa abordagem seja simples, ela possuí várias limitações, pois exige o esforço da inserção manual de informações e, além disso, as informações inseridas são subjetivas (pessoas diferentes têm diferentes percepções da mesma imagem) [Zhang & Su, 2002]. A segunda abordagem é a conhecida por recuperação de imagens por conteúdo (CBIR), onde o conteúdo visual das imagens é usado para recuperá-las. A busca de imagens por conteúdo envolve várias etapas. Inicialmente as imagens são processadas e as características visuais das mesmas são extraídas. A similaridade entre as imagens é avaliada para a construção de índices. Para a execução da consulta, as características da imagem de consulta são extraídas e comparadas com as características das demais imagens da base.

Neste capítulo são abordadas as principais tecnologias usadas nos sistemas PACS, CAD e CBIR. Além disso, são discutidos os conceitos envolvidos em cada etapa da busca de imagens por conteúdo, exceto a etapa de processamento e extração de características, que é detalhada no capítulo 2.

3.2 Sistemas PACS

A tecnologia PACS pode ser utilizada para visualizar e propiciar a análise de imagens, mesmo sem utilizar a impressão de imagens em filme (*filmless*). Um hospital que use um ambiente *filmless* deve possuir um ambiente de rede amplo e integrado, no qual o filme foi completamente, ou em grande parte, substituído por sistemas eletrônicos que coletam as imagens, efetuam seu arquivamento, disponibilizando-as e apresentando-as em dispositivos de saída. Um sistema PACS deve executar as seguintes funções utilizando a tecnologia digital [Rosa, 2007]:

- Aquisição de imagem;
- Comunicação de imagens (transferência tanto entre dispositivos quanto entre lugares físicos);
- Armazenamento de imagens;
- Exibição de imagens;
- Processamento de imagens.

Os sistemas PACS em conjunto com os Sistemas de Informação Hospitalar (SIH) e os Sistemas de Informação Radiológica (SIR) formam a base para propiciar um ambiente *filmless*. A figura 3.1 apresenta a organização geral de um sistema PACS, onde os dispositivos de aquisição

como tomógrafo computadorizado, tomógrafo de ressonância magnética e equipamentos de radiologia computadorizada são conectados à rede que interliga os servidores de dados e imagens e os monitores de visualização.

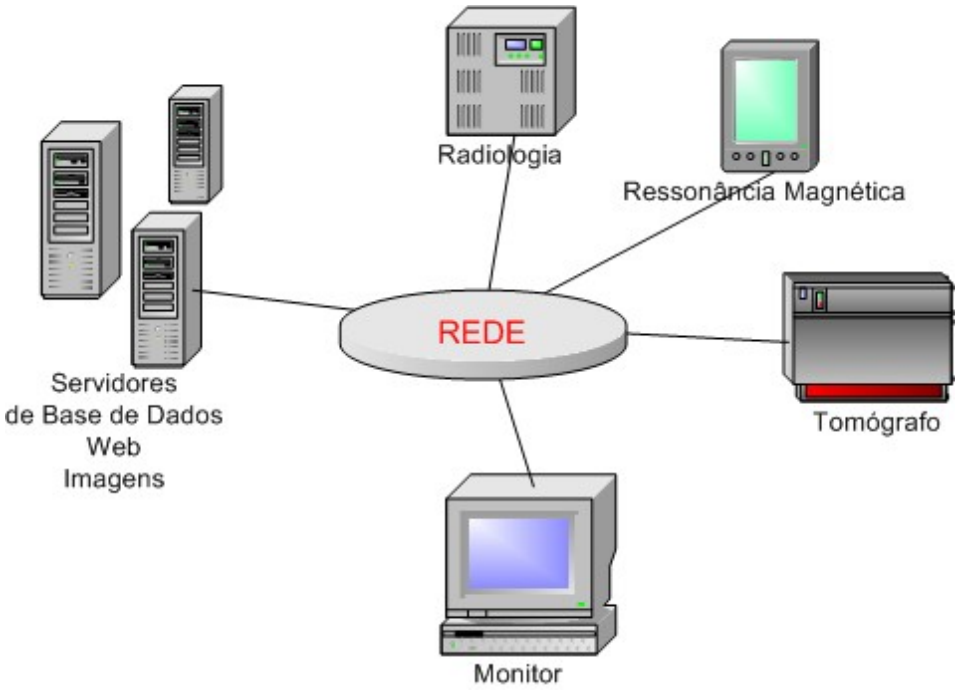


Figura 3.1: Arquitetura de um sistema PACS [Rosa, 2002]

.

Infelizmente, os poucos sistemas PACS comerciais oferecidos atualmente são extremamente caros e não contemplam todas as necessidades do centro médico. Dessa forma, o estado da arte na área é o desenvolvimento de soluções locais e muitas vezes restritas [Rosa et al., 2002]. Cerca de 70% da carga de trabalho de departamentos de radiologia utilizam a tecnologia convencional de filmes, muito embora tenha se desenvolvido as tecnologias que obtenham diretamente a imagem no formato digital, como ressonância magnética (*magnetic resonance* - MR), tomografia computadorizada (*computed tomography* - CT), ultra-som e angiografia digital [Siegel & Kolodner, 1999].

A maioria dos equipamentos para modalidades digitais, tais como tomografia computadorizada, ressonância magnética e radiografia digital direta (*Direct Digital Radiography* - DR), usa a interface padrão, o protocolo de comunicação e formato de imagem conhecido como DICOM (*Digital Imaging and Communications in Medicine*) [Rangayyan, 2005]. O formato DICOM associa as imagens com informações textuais, e é utilizado por diversas modalidades de equipamentos de imagens médicas.

Uma vez que as imagens são obtidas, elas devem ser armazenadas para uso futuro. O armazenamento das imagens tem se dividido, tradicionalmente, em curto período (*short-term*), o qual inclui um meio de armazenamento magnético e local, e longo período (*long-term*), o qual

envolve meios de armazenamento óticos entre outros. Nos PACS, tipicamente a maior parte das imagens ficam armazenadas em dispositivos de longo período, que são mais lentos do que os dispositivos de curto período [Siegel & Kolodner, 1999]. O gerenciamento desses armazenamentos geralmente é feito através de um sistema gerenciador de base de dados que mantém o controle da localização e movimentação das imagens.

3.3 Sistemas CAD

Os sistemas CAD (*Computer Aided Diagnosis*) são sistemas que usam estratégias e o poder computacional para fornecer ao radiologista uma segunda opinião sobre o diagnóstico de uma determinada imagem médica. Uma série de estudos realizados na Universidade de Chicago [Quek et al., 2003] [MacMahon et al., 1999] [Kobayashi & Doi, 1999] constataram que a segunda opinião fornecida pelos sistemas CAD é importante para aumentar a precisão dos diagnósticos e a consistência da interpretação da imagem radiológica, mediante ao uso da resposta do computador como referência. A figura 3.2 apresenta curvas ROC (*Receiver Operating Characteristic*) [Doi, 2005] comparando a precisão na distinção entre nódulos malignos e benignos de pulmão feita por radiologistas com e sem o auxílio de sistemas CAD. Em uma curva ROC, quanto maior é a área abaixo da curva, melhor é o método. Assim, através das curvas apresentadas na figura 3.2 é possível perceber que o uso de sistemas CAD melhorou a precisão da distinção entre nódulos malignos e benignos de pulmão. Na seção 3.3.1 são detalhados os conceitos associados a curvas ROC.

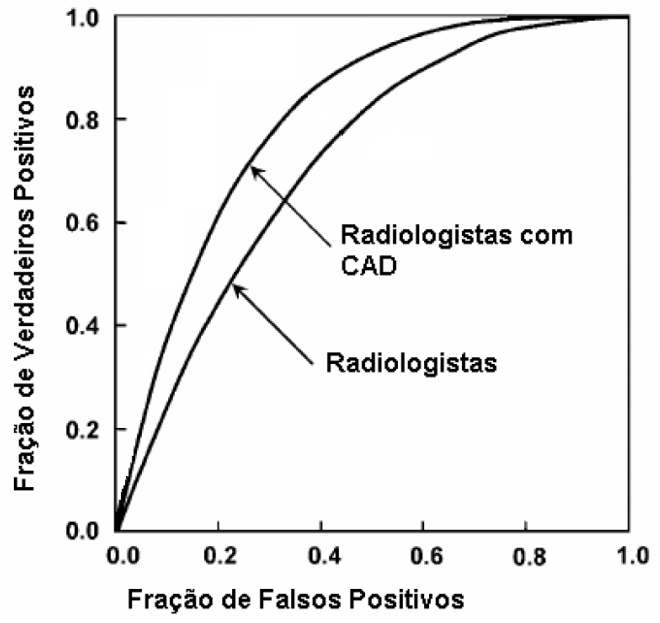


Figura 3.2: Curvas ROC para a distinção entre nódulos malignos e benignos com e sem o uso de sistemas CAD [Doi, 2005].

O laudo feito por um ser humano pode ser afetado por uma série de fatores, como a presença de ruídos na imagem, a semelhança de características entre exames normais e anormais, o estado emocional e o cansaço dos radiologistas. O uso da segunda opinião obtida através de um sistema CAD auxilia na redução de erros nos laudos feitos. Por exemplo, tem-se que a mamografia é o melhor método para detectar câncer de mama, no entanto, uma faixa de 10% a 30% das mulheres que possuem câncer de mama, tem laudos negativos na mamografia, desses falsos negativos, 66% representaram falhas nos laudos onde a detecção de câncer poderia ser obtida diretamente pela análise da imagem [Giger, 2000]. Quando o radiologista detecta a presença de um tumor, ele tenta distinguir pelas características o tipo da lesão, recomendando a biopsia em caso de suspeita de ser maligno. Embora existam regras gerais que diferenciem lesões malignas ou benignas, existe uma grande variedade na interpretação das mesmas. Apenas 10% a 20% dos tecidos submetidos a um procedimento cirúrgico de biopsia são confirmados como tumor maligno. Estudos mostram que essa grande faixa de intervenções cirúrgicas desnecessárias para a distinção entre os tumores malignos e benignos tem sido diminuída com o uso de CAD [Giger, 2000]. Em [Freer & Ulissey, 2001] foi apresentado um estudo que mostra que o uso de sistemas CAD promoveu um ganho de 20% na detecção, em estágios iniciais, de tumores de mama.

As principais tecnologias usadas em sistemas CAD são [Doi, 2005]:

- Processamento de imagens para a detecção e extração de anormalidades;
- Determinação e quantização de características para detectar anormalidade;
- Mineração de imagens para a classificação de regiões normais ou anormais;
- Avaliação do ganho de desempenho de radiologistas através do uso de curvas ROC;
- Avaliação da busca de imagens similares a uma imagem com um determinado tipo de lesão.

Os sistemas CAD podem ser usados em todas as modalidades de imagens médicas, incluindo radiografias, CT, MR, ultra-som, e em todas as partes do corpo humano como esqueleto, tórax, coração e abdômen. Entretanto, a maior parte dos sistemas CAD desenvolvidos até agora tratam da detecção de tumores em mamografias [Baeg & Kehtarnavaz, 2000] [Comer et al., 1996] [Jiang et al., 2001] [Hadjiiski et al., 2004] [Quek et al., 2003] e nódulos em pulmão através de radiografia e CT do tórax [Kobayashi & Doi, 1999] [Kobayashi et al., 1996] [MacMahon et al., 1999]. Assim, foi explorada apenas uma pequena parcela de toda a gama de aplicações dos sistemas CAD, por isso, uma grande evolução dos sistemas CAD no futuro é esperada. Existem basicamente dois tipos de auxílios prestados pelos sistemas CAD [Marques, 2001]:

- *Auxílio à detecção de lesões*: localização de padrões anormais através da varredura da imagem pelo computador (por exemplo, detecção de nódulos pulmonares em imagens do tórax);

- *Auxílio ao diagnóstico*: quantificação das características das imagens e sua classificação em padrões normais ou anormais (por exemplo, uso das características de forma do tecido para associá-lo a um tipo de tumor).

A detecção automática de lesões envolve a localização pelo computador de regiões contendo padrões radiológicos que podem indicar uma lesão, no entanto, a análise e classificação da lesão devem ser feitas por um radiologista. A análise computacional automática utiliza características visuais automaticamente extraídas das imagens. No entanto, é necessário que previamente sejam definidos, com o auxílio dos especialistas, quais são os atributos clinicamente significativos [Marques, 2001].

A análise de mamografias por sistemas CAD é muito importante. A Sociedade Americana de Câncer recomenda que as mulheres com idade entre 35 e 39 anos façam um exame de mamografia básico de prevenção. A partir dos 40 anos de idade, as mulheres devem fazer esse exame a cada 2 anos e, a partir dos 50 anos ou mais, a mamografia deve ser feita uma vez por ano. Esses exames auxiliam o médico a diagnosticar o câncer de mama em uma fase inicial onde existe uma maior chance de cura. Segundo dados da Sociedade Americana de Câncer, uma a cada oito mulheres nos Estados Unidos tem câncer de mama durante sua vida. A mamografia permite o diagnóstico de alguns tipos de câncer de mama de 1 a 2 anos antes do médico ou do paciente poder detectá-lo no auto-exame. Segundo o Instituto Nacional do Câncer (INCA), diferenciar o tecido normal do anormal da mama é uma tarefa complexa, pois a mama tem um tecido dinâmico que se modifica durante a vida das mulheres. A detecção de nódulos é mais difícil do que a detecção de microcalcificações em mamografias, isso porque os nódulos se assemelham com o tecido que os envolve [Marques, 2001]. Fatores físicos, equipamento e sobreposição de elementos interferem na qualidade da imagem gerada fazendo com que seu diagnóstico seja prejudicado.

Em se tratando de radiografia de tórax, existe uma série de patologias possíveis. O câncer do pulmão tem sido indicado como um dos principais causadores de morte de homens e mulheres por câncer no mundo. A detecção precoce dessa anomalia é importante para aumentar a taxa de sobrevida dos pacientes. A radiografia do tórax é considerada um dos exames mais práticos para auxiliar diagnósticos, porém a taxa de erros dos diagnósticos efetuados por radiologistas chega a 30% e estudos mostram que usando os sistemas CAD a taxa de acerto chega a 80% [Kobayashi & Doi, 1999] [Doi, 2007].

3.3.1 Medindo o Desempenho de Sistemas CAD

Um dos métodos mais utilizados para medir a eficácia de um sistema CAD é a curva ROC. Os conceitos necessários para construir e analisar uma curva ROC são detalhados a seguir.

Para facilitar o entendimento, considere que um teste seja um diagnóstico feito por um sistema CAD ou por um radiologista com o intuito de determinar a existência de uma determinada anomalia. A idéia básica de interpretar o resultado de um teste é *calcular a probabilidade do paciente ter uma doença, dado um diagnóstico*. Para isso, é importante entender o conceito de verdadeiros positivos (*true positives* - TP), verdadeiros negativos (*true negatives*- TN), falsos positivos (*false positives* -FP) e falsos negativos (*false negatives* -FN) que é ilustrado na tabela 3.1. Um teste é positivo quando ele indica a presença de uma anomalia existente. Um teste é negativo quando ele indica a ausência real da mesma. Uma anomalia é verdadeira quando ela existe, falsa quando não existe.

Tabela 3.1: Conceito de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

	Presença de Anomalia A+	**Ausência de Anomalia A-**
Teste Positivo (T+)	Verdadeiros Positivos (TP)	Falsos Positivos (FP)
Teste Negativo (T-)	Falsos Negativos (FN)	Verdadeiros Negativos (TN)

As duas medidas mais importantes usadas na análise dos resultados de um teste são a sensibilidade e a especificidade.

A *sensibilidade* é a proporção dos pacientes com anomalias cujo teste foi positivo. Na notação de probabilidade, a sensibilidade é:

$$P(T+|A+) = \frac{TP}{TP+FN} \tag{3.1}$$

A *especificidade* indica a porcentagem dos pacientes que não possuem anomalias cujo teste foi negativo. Na notação de probabilidade, a especificidade é:

$$P(T-|A-) = \frac{TN}{TN+FP} \tag{3.2}$$

A sensibilidade e a especificidade descrevem quão bem o teste se comporta em discriminar pacientes com ou sem a anomalia. Na análise por curva ROC, os resultados dos testes podem ser modelados como uma variável que assume dois valores: *normal*(ausência de anomalia) e *anormal* (presença de anomalia).

Considere o gráfico da Figura 3.3 que ilustra o número de casos normais e anormais versus o valor de um atributo T. Na maioria dos casos, as distribuições se sobrepõem e o resultado do teste não distingue os casos normais dos anormais com 100% de precisão. A área de sobreposição indica a região onde o teste não consegue distinguir os casos normais dos anormais. Um

valor de limiar para o atributo T deve ser escolhido para separar os casos normais dos anormais. Esse valor determina o número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, envolvendo sempre um compromisso entre a sensibilidade e a especificidade. Assim, deve ser escolhido um valor de limiar que minimize os custos associados aos resultados falsos negativos e falsos positivos.

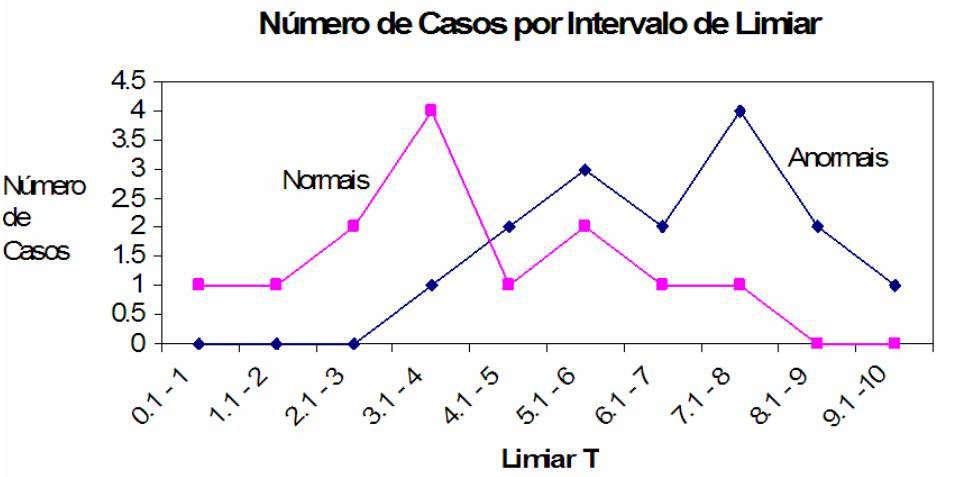


Figura 3.3: Distribuição de casos normais e anormais por intervalos do limiar T.

Construção de curvas ROC

Para ilustrar a construção de uma curva ROC, é usada a tabela 3.2 que mostra o número de casos normais e anormais encontrados para cada intervalo do atributo T. Para compreender melhor os dados da tabela 3.2, considere, por exemplo, que T seja a temperatura dos pacientes e que o valor de T seja usado para determinar se o estado do paciente é normal ou anormal. A primeira linha da tabela 3.2 indica que apenas um paciente teve temperatura no intervalo [0,1] e o estado desse paciente é normal. Já, a sexta linha da tabela 3.2 indica que 5 pacientes tiveram temperatura no intervalo [5.1,6], sendo que o estado de 2 desses pacientes é normal e o estado de 3 desses pacientes é anormal.

Os dados da tabela 3.2 são usados para determinar os valores de sensibilidade e especificidade calculados de acordo com a escolha do valor do limiar T. Considere que seja escolhido o valor de limiar T=5. Os casos com $T \leq 5$ serão apontados como normais (negativos) e os casos com $T > 5$ serão apontados como anormais (positivos). De um total de 13 casos comprovadamente normais, 9 foram indicados como normais através da escolha do limiar T=5 e, de um total de 15 casos comprovadamente anormais, 12 casos foram indicados como anormais. Através desses dados é possível calcular a sensibilidade e a especificidade para esse valor de limiar. Para o valor limiar $T = 5$, a sensibilidade é dada por $TP/(TP+FN) = 12/(12+3) = 0.8$. Já, a especificidade é dada por $TN/(TN+FP) = 9/(9+4) = 0.69$. A fração de falsos positivos é calculada a partir da especificidade:

Tabela 3.2: Número de casos normais e anormais em função dos valores de *T*.

T	**Casos Anormais**	**Casos Normais**
0.1-1	0	1
1.1-2	0	1
2.1-3	0	2
3.1-4	1	4
4.1-5	2	1
5.1-6	3	2
6.1-7	2	1
7.1-8	4	1
8.1-9	2	0
9.1-10	1	0
total	**15**	**13**

fração de falsos positivos = 1 - especificidade

A curva ROC é o gráfico da sensibilidade (fração de verdadeiros positivos) versus 1-especificidade (fração de falsos positivos), onde os pontos do gráfico são calculados a partir da escolha de diferentes limiares. A tabela 3.3 mostra os valores de sensibilidade, especificidade e a fração de falsos positivos calculados a partir da escolha dos limiares 5 ,7 e 9 para *T*. Observe através da tabela 3.3 que quanto maior a sensibilidade menor a especificidade e vice-versa. A partir dos dados da tabela 3.3 tem-se 3 pontos para traçar uma curva ROC. Essa curva é ilustrada na figura 3.4.

Tabela 3.3: Valores de sensibilidade, especificidade e fração de falsos positivos calculados a partir dos valores de limiar 5, 7 e 9.

T	Intervalo	Anormais	Normais	Sensibilidade	Especificidade
5	≤ 5	3	9	0.8	0.69
	> 5	12	4		
7	≤ 7	8	12	0.47	0.92
	> 7	7	1		
9	≤ 9	14	13	0.07	1
	> 9	1	0		

A partir da curva ROC é possível medir a precisão do método através da área abaixo da curva. A área abaixo da curva varia de 0.5 (pior caso, comportamento aleatório) até 1.0 (melhor caso, discriminação perfeita). Assim, quanto mais perto da borda superior estiver a curva, melhor é o método; quanto mais a curva se aproxima da diagonal do gráfico, pior é o método.

3.4 Sistemas CBIR

Os sistemas de recuperação de imagens por conteúdo (CBIR) abordam tecnologias e métodos voltados para organização de grandes repositórios de imagens digitais por meio do conteúdo vi-

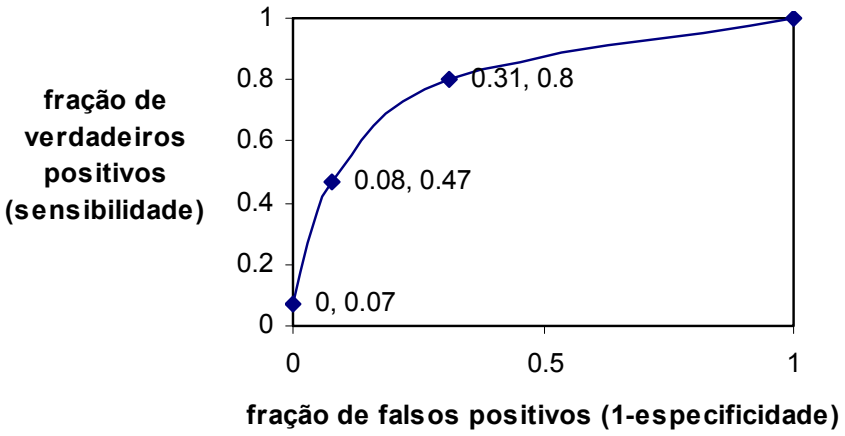


Figura 3.4: Curva ROC obtida a partir dos dados da tabela 3.3.

sual das mesmas. Várias áreas da computação contribuem para o desenvolvimento dos sistemas CBIR, dentre elas destacam-se: base de dados, processamento de imagens, visão computacional e interação usuário-computador. Os principais tópicos referentes a construção de sistemas CBIR para a análise de imagens estão relacionados com:

- *Extração de características de imagens*: Extração automática de características visuais de baixo nível (cor, textura e forma) que possam representar adequadamente a imagem de acordo com seu significado e interpretação humana de alto-nível. A definição de características visuais que representam adequadamente o conteúdo semântico da imagem tem sido um dos grandes desafios dos pesquisadores da área de CBIR. Essa inconsistência entre a informação de baixo nível automaticamente extraída das imagens e a interpretação humana de alto nível é chamada de "gap semântico" [Deserno et al., 2008]. Para lidar com o "gap semântico", algoritmos poderosos para extração de características para determinados tipos de imagens têm sido desenvolvidos [Silva, 2007] [Balan, 2007] além de métodos seletores de características, que buscam encontrar as características mais importantes para discriminar as imagens em um dado vetor de características [Ribeiro et al., 2005a];

- *Definição de funções de distância*: É a função de distância que indica o nível de semelhança entre pares de imagens, quando aplicada a vetores de características delas obtidos. No entanto, qual função de distância é mais adequada para utilizar em qual contexto? O trabalho desenvolvido por [Bugatti, 2008] vem a tentar responder essa questão, onde o desempenho de diferentes funções de distâncias é comparado utilizando diferentes extratores de características. Além disso, trabalhos têm sido desenvolvidos para determinar qual é a melhor combinação de características e funções de distância que deve ser utilizada para cada tipo de imagem médica [Silva, 2008];

- *Indexação das características*: Para tornar as operações de consultas eficientes, de maneira que o resultado da consulta seja retornado em um tempo aceitável é necessário criar índices de imagens e organizá-los em estruturas de dados apropriadas. Vários trabalhos para indexação de dados métricos [Traina Jr. et al., 1997] [Traina Jr. et al., 2007] [Wichert, 2008] têm sido desenvolvidos e aplicados com sucesso em sistemas CBIR.

- *Interface com o usuário*: Sistemas CBIR são inerentemente interativos. A interação do sistema com o usuário é geralmente feita através de uma interface que permita a seleção de tipos de consultas e imagens e de um navegador que possibilite visualizar os resultados. Além das funcionalidades tradicionais, sistemas atuais também permitem a seleção direta de características e funções de distância, além de refazer a consulta utilizando informações de realimentação de relevância [Rosa, 2007].

A Figura 3.5 apresenta uma visão geral de um sistema CBIR e seus principais componentes.

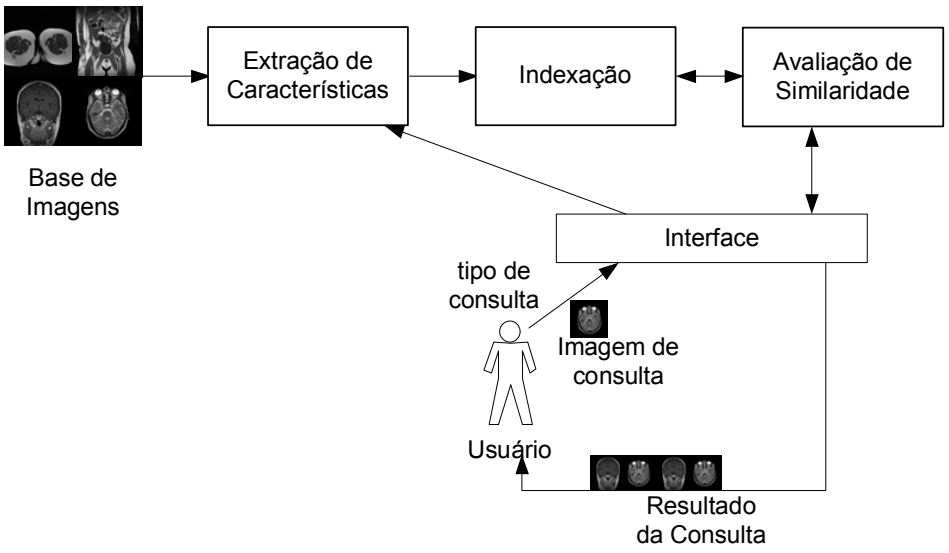


Figura 3.5: Visão geral de um sistema CBIR.

O procedimento de realimentação de relevância, que permite a interação do usuário para o refinamento da consultas em sistemas CBIR, é uma alternativa que vem sendo utilizada com sucesso para vencer o "gap semântico" existente entre a representação das imagens, através de vetores de características, e seu significado semântico [Marques et al., 2008]. A realimentação de relevância é um procedimento que permite reformular a consulta, adicionando novos centros de consulta, ajustando a função de distância utilizada, e ponderando os vetores de características de acordo com a realimentação fornecida pelo usuário, melhorando os resultados das consultas nas próximas interações. A reformulação da consulta pode ser feita empregando algoritmos genéticos [Ferreira et al., 2008], técnicas estatísticas [Ribeiro et al., 2006c] e lógica nebulosa [Krishnapuram et al., 2004], dentre outras técnicas.

Dentre os sistemas CBIR comerciais mais conhecidos pode-se citar o QBIC [Flickner & alli, 1995] da IBM e o AMORE [Mukherjea et al., 1999] da Nec. Dentre os sistemas acadêmicos desenvolvidos mais conhecidos estão o Photobook [Pentland et al., 1996], o Netra [Ma et al., 1997], o GIFT *GNU Image Finding Tool* (http://www.gnu.org/software/gift) e o IRMA *http://irma-project.org/*. Entre os sistemas desenvolvidos mais recentemente estão o SiREN [Barioni, 2006], o BIRAM [Moreno & Furuie, 2006] e o cbPACS [Rosa, 2007], que integra as funcionalidade de busca por conteúdo em um PACs. Nesses sistemas, o objeto de consulta pode ou não fazer parte do conjunto de resposta dependendo da implementação.

3.4.1 Indexação

A presença de um esquema de indexação que permita uma rápida comparação entre as múltiplas características de uma imagem é muito importante para melhorar o desempenho da mineração de imagens. Várias técnicas de indexação, ou métodos de acesso, foram propostos na literatura. Dependendo do modelo utilizado para representar os objetos, existem os seguintes tipos de métodos de acesso:

- *Métodos de Acesso Espaciais (MAEs):* voltados para o modelo no qual os objetos são representados por vetores em um espaço multidimensional. Exemplos de MAEs dinâmicos são: a R-tree [Guttman, 1984], a TV-Tree [Lin et al., 1994], a SR-Tree [Katayama & Satoh, 1997], a KDB-tree e suas variantes [Lin & Chen, 2008];

- *Métodos de Acesso Métricos (MAMs)*: voltados para o modelo no qual apenas a distância entre os objetos é levada em consideração. Exemplos conhecidos de MAMs dinâmicos são: a M-Tree [Ciaccia et al., 1998], a Slim-Tree [Traina Jr. et al., 2000a], a DF-tree [Traina Jr. et al., 2002] e a família OMNI [Traina Jr. et al., 2007].

Em geral, os MAMs consistem em selecionar um ou mais objetos e colocá-los como representativos ou pivôs do conjunto. Quando um objeto é inserido na estrutura, as distâncias entre ele e seus representativos são calculadas e armazenadas. Essas distâncias são usadas durante as consultas no descarte de objetos. Os MAMs suportam naturalmente consultas por similaridade. Dessa forma, tem-se mostrado bastante apropriado utilizar MAMs para indexar os atributos que foram extraídos das imagens, suportando a busca por similaridade.

A função de distância usada nos MAMs deve ser obrigatoriamente métrica. Considerando os objetos x, y e z, pertencentes ao domínio D, uma função de distância $d(x,y)$ deve satisfazer as seguintes propriedades:

- Simetria: $d(x,y) = d(y,x)$;

- Não negatividade: $0 \leq d(x,y) < \infty$

- Auto-similaridade: $d(x,x) = 0$;
- Desigualdade triangular: $d(x,y) \leq d(x,z) + d(z,y)$.

A propriedade de desigualdade triangular é bastante útil para o processo de indexação, pois permite realizar podas dos objetos no espaço de busca, tornando o processo de recuperação da informação mais rápido. Quando os objetos a serem comparados são vetores n-dimensionais com n fixo, tem-se um caso particular de espaço métrico que é o espaço vetorial.

3.4.2 Tipos de Consultas por similaridade

Critérios baseados em igualdade são inadequados para imagens (e muitos tipos de dados complexos), pois não há sentido em realizar consultas como: "obtenha as imagens com tumores no cérebro iguais à imagem em estudo". Dificilmente (provavelmente nunca) as imagens de dois tumores serão exatamente iguais, mesmo que os tumores tenham a mesma classificação. O critério mais adequado para os casos desse tipo é o de similaridade. Assim, a consulta anterior faria mais sentido se definida como: "obtenha as imagens com tumores no cérebro semelhantes à imagem em estudo". Uma consulta por similaridade, de acordo com a sua abrangência ou cobertura, é classificada em consulta por abrangência ou consulta por vizinhança. Esses dois tipos de consultas são detalhados a seguir.

Na consulta por abrangência (*range query*), são fornecidos um objeto de referência Q e um raio de cobertura r. O conjunto resposta R_{rq} inclui todos os elementos S da base que se encontram a uma distância menor ou igual a r do elemento Q. Ou seja:

$R_{rq} = \{S \mid d(S,Q) \leq r\}$

A Figura 3.6 ilustra um exemplo de consulta por raio de abrangência no domínio bi-dimensional, onde o conjunto de resposta contém oito elementos. A função de distância utilizada neste caso é a função euclidiana.

Na consulta aos k-vizinhos mais próximos (kNN) são fornecidos um objeto de referência Q e um número inteiro k referente ao número de elementos mais próximos do elemento Q que se deseja obter como conjunto de resposta R_{kNN}. Formalmente tem-se:

$R_{kNN} = \{S \mid \forall\, P \in \{\Omega - R_{kNN}\}, d(Q,S) \leq d(Q,P),\ |R_{kNN}| = K\}$,

onde Ω representa o conjunto de todos os elementos.

A figura 3.7 ilustra um exemplo de consulta do tipo kNN no domínio bi-dimensional, onde o conjunto de resposta contém seis elementos.

3.4.3 Funções de distância

A tarefa de avaliação de similaridade entre elementos de um repositório de imagens é realizada através da aplicação de uma função de distância. Nos casos de estudo desta tese, foram

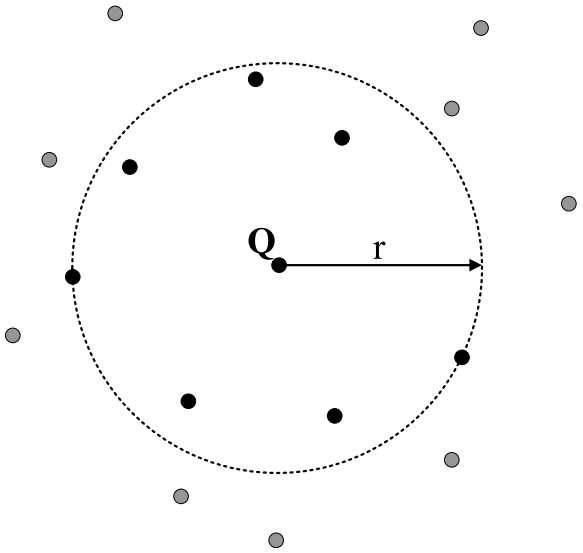


Figura 3.6: Exemplo de uma consulta por raio de abrangência onde o conjunto de resposta contém 8 elementos.

utilizadas as funções de distância métricas, pois elas tendem a refletir a percepção humana de similaridade.

Aqui são descritas as funções de distância utilizadas nos experimentos desta tese: família de distâncias de Minkowski (família L_p), a distância de Mahalanobis, a divergência de Jeffrey, a distância χ^2 e a distância Camberra. Considerando, os dois vetores de N características $R = [R_0, R_1, ..., R_{N-1}]$ e $S = [S_0, S_1, ..., S_{N-1}]$, essas funções de distância são detalhadas a seguir.

Funções de distâncias de Minkowski (família L_p)

A família de distâncias proposta por Minkowski é conhecida como família de distâncias L_p e é apresentada na equação 3.3.

$$L_p(R,S) = (\sum_{i=0}^{N-1} |R_i - S_i|^p)^{\frac{1}{p}} \tag{3.3}$$

Três funções de distância bem conhecidas fazem parte da família L_p. Quando $p = 1$, tem-se a distância L_1 (equação 3.4) que corresponde à distância *Manhattan*, ou *city block*. Quando $p = 2$ tem-se a distância Euclidiana L_2 (equação 3.5) , que é uma das funções de distância mais conhecidas e utilizadas. Fazendo $p \to \infty$ tem-se a distância L_∞ ou $L_{infinity}$ definida na equação 3.6. A Figura 3.8 ilustra as configurações de um conjunto de pontos eqüidistantes considerando as distâncias L_1, L_2 e L_∞ em um espaço bi-dimensional.

$$L_1(R,S) = \sum_{i=0}^{N-1} |R_i - S_i| \tag{3.4}$$

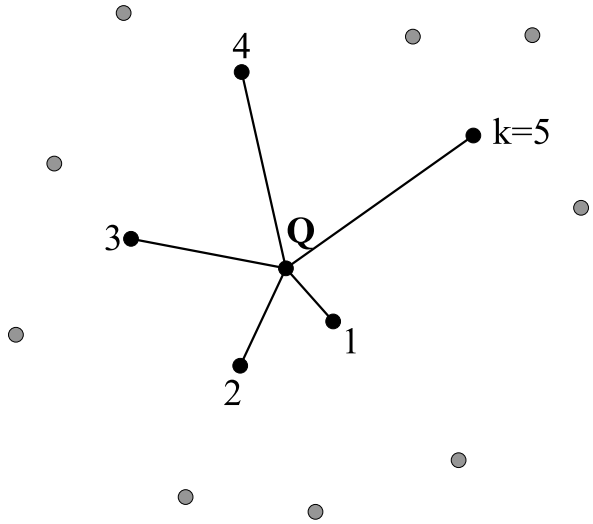


Figura 3.7: Exemplo de uma consulta do tipo *k*NN onde o conjunto de resposta contém 6 elementos.

$$L_2(R,S) = (\sum_{i=0}^{N-1} (R_i - S_i)^2)^{\frac{1}{2}} \tag{3.5}$$

$$L_\infty(R,S) = \max_{0 \le i < N} [\ |R_i - S_i|\ ] \tag{3.6}$$

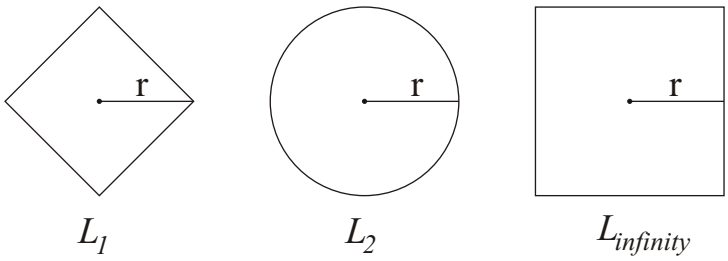


Figura 3.8: Configurações de um conjunto de pontos eqüidistantes para as distâncias L_1, L_2 e $L_{infinity}$ em um espaço bi-dimensional.

Na família L_p, a L_2 é a única distância que define um lugar onde todos os pontos encontram-se geometricamente eqüidistantes do ponto de referência. A L_1 reduz a distância entre os objetos que concentram suas dissimilaridades em relação a apenas um dos atributos considerados. Já, a L_∞ reduz a distância quando as diferenças são melhor distribuídas entre os atributos. Existe uma variação da família *Lp* que é a distância de Minkowski ponderada, na qual atribui-se pesos para cada característica do vetor que representa a imagem. Esses pesos são valores atribuídos a um vetor, chamado de vetor de ponderação. A utilização de um vetor de ponderação visa tratar diferentes influências das características sobre a percepção de similaridade entre as imagens em

diferentes contextos. A distância de Minkowski ponderada é definida formalmente na equação 3.7, onde w é o vetor de ponderação $w = (w_1, w_2, ..., w_n)$.

$$d_{L_p}(R,S) = \sqrt[p]{\sum_{i=1}^{n} w_i (R_i - S_i)^p} \tag{3.7}$$

Distância quadrática e distância de Mahalanobis

A função de distância quadrática leva em consideração não apenas a correspondência existente entre as características de mesmo índice no vetor. Também são levadas em consideração as informações sobre a similaridade de elementos de índices diferentes. A distância quadrática é definida como:

$$d_{quad}(R,S) = \sqrt{(R-S)^T A (R-S)} \tag{3.8}$$

onde A é uma matriz de dimensão $N \times N$ de elementos $a_{i,j}$ que correspondem ao coeficiente de afinidade entre os elementos de índices i e j. O valor de $a_{i,j}$ é dado por:

$$a_{i,j} = 1 - \frac{d_{i,j}}{\max[d_{i,j}]} \text{onde} \qquad d_{i,j} = |R_i - S_j|$$

A distância Mahalanobis é um caso especial da distância quadrática, onde a matriz de transformação A corresponde à matriz inversa da matriz de covariância, obtida a partir de um conjunto de treino de vetores de características. Para se alcançar a definição desta função é preciso considerar um vetor de variáveis aleatórias $X = [X_0, X_1, ..., X_{N-1}]$ que assume os valores das características dos vetores que constituem um conjunto de treino estabelecido. A matriz de covariância V é dada por $V = [\sigma^2_{i,j}]$, onde $\sigma^2_{i,j} = E[X_i X_j] - E[X_i]E[X_j]$. Assim, a distância Mahalanobis entre dois vetores R e S é definida como:

$$d_{mah} = \sqrt{(R-S)^T V^{-1} (R-S)} \tag{3.9}$$

No caso especial onde as variáveis $X_0, X_1, ..., X_{N-1}$ são estatisticamente independentes, a matriz de covariância V é a matriz diagonal:

$$V = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}$$

Neste caso a distância Mahalanobis fica reduzida à seguinte forma:

$$d_{mah-i}(R,S) = \sum_{i=0}^{N-1} \frac{(R_i - S_i)^2}{\sigma_i^2} \tag{3.10}$$

A equação 3.10, corresponde a uma função de distância L_2 ponderada pela variância. Para as características que possuem pequena variância dentro do conjunto de treino é atribuído um peso maior, ao passo que para as características de maior variância o peso é menor.

Divergência de Jeffrey

A divergência de Jeffrey é uma distância utilizada por ser bastante tolerante a ruídos [Rubner & Tomasi, 2001]. Ela é dada pela seguinte equação, onde $m_i = \frac{R_i+S_i}{2}$:

$$d_J(R,S) = \sum_{i=1}^{n} \left(R_i \log \frac{R_i}{m_i} + S_i \log \frac{S_i}{m_i} \right) \tag{3.11}$$

Distância χ^2

A distância χ^2 enfatiza discrepâncias entre dois vetores de características. Ela é dada pela seguinte equação:

$$d_{\chi^2}(R,S) = \sum_{i=1}^{n} \frac{(R_i - m_i)^2}{m_i}, \text{ where } m_i = \frac{R_i + S_i}{2} \tag{3.12}$$

Distância Camberra

A distância Camberra é semelhante a distância de Manhattan com relação à restrição do espaço de retorno, e ela é dada pela seguinte equação:

$$d_C(R,S) = \sum_{i=1}^{n} \frac{|R_i - S_i|}{|R_i| + |S_i|} \tag{3.13}$$

3.4.4 Avaliação de eficiência

Uma das questões envolvendo sistemas CBIR é como medir a qualidade do resultado obtido em uma operação de consulta. Um dos instrumentos mais utilizados para avaliar a eficácia dos sistemas de busca são os gráficos de precisão e revocação (*precision vs. recall* - P&R) [Baeza-Yates & Ribeiro-Neto, 1999]. A seguir são apresentados detalhes da construção de tais gráficos.

Para uma dada consulta, considere que *Re* seja o número total de itens relevantes existentes na base, Re_R o número total de itens relevantes recuperados e I_R o total de itens recuperados.

Revocação é a fração do conjunto de elementos relevantes (*Re*) que foram recuperados na consulta.

$$\text{Revocação} = \frac{Re_R}{Re} \tag{3.14}$$

Precisão é a fração do conjunto de elementos recuperados (I_R) que são relevantes.

$$\text{Precisão} = \frac{Re_R}{I_R} \tag{3.15}$$

Para a construção do gráfico de P&R, o primeiro passo é ordenar os elementos recuperados da base de acordo com sua distância em relação ao objeto de consulta. Como um exemplo prático [Baeza-Yates & Ribeiro-Neto, 1999], considere que para uma determinada consulta q existe na base um conjunto Rq de 10 elementos relevantes, composto da seguinte maneira:

$Rq = \{e_5, e_{13}, e_{17}, e_{20}, e_{31}, e_{36}, e_{42}, e_{47}, e_{55}, e_{61}\}$

Considere que um algoritmo de busca tenha retornado um conjunto de elementos Iq, referentes à consulta q, cujos elementos e suas respectivas relevâncias são dados por:

1. e_{42} •	2. e_{54}
3. e_{17} •	4. e_{15}
5. e_2	6. e_5 •
7. e_{25}	8. e_{27}
9. e_{13} •	10. e_{54}
11. e_{55} •	12. e_{67}

Os elementos que são relevantes à consulta q estão marcados com •. Examinando o conjunto Iq dos elementos recuperados, verifica-se que o primeiro elemento (e_{42}) é um dos elementos relevantes à consulta. Neste caso, o valor de precisão é de 100%, pois de todos os elementos analisados (apenas o primeiro até aqui), são relevantes à consulta. Neste ponto, o valor de revocação é 10%, pois um elemento relevante, dentro de um conjunto de dez elementos, foi recuperado até esse ponto. O próximo elemento relevante da lista é o terceiro elemento (e_{17}). Para este elemento o valor de precisão é de aproximadamente 66% (dois elementos relevantes em três verificados) e o valor de revocação é de 20% (dois entre dez elementos relevantes). A análise prossegue desta maneira até que todos os elementos relevantes sejam verificados. Os valores de precisão e revocação são traçados no gráfico de P&R, conforme o exemplo ilustrado na figura 3.9.

Para uma avaliação confiável dos resultados obtidos por um determinado sistema de recuperação, é necessário que diversas operações de consultas sejam realizadas e consideradas na avaliação. Para isso, é preciso construir uma curva de precisão e revocação que represente a média dos desempenhos das diversas consultas realizadas. Isto é feito, geralmente, calculando-se valores de precisão para escalas determinadas de revocação, como a cada intervalo de 10% de revocação.

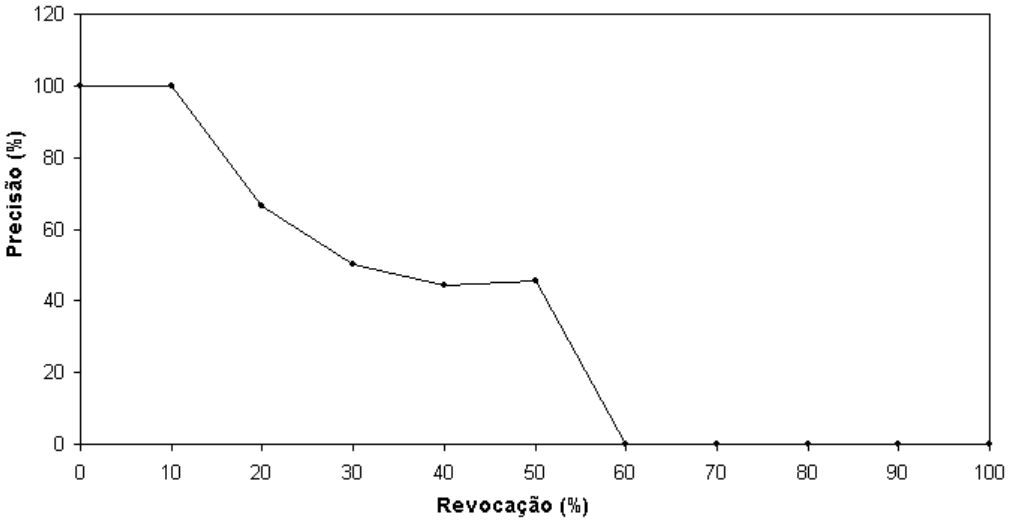


Figura 3.9: Exemplo de um gráfico de medidas P&R para uma operação de busca [Baeza-Yates & Ribeiro-Neto, 1999].

A avaliação do desempenho é realizada observando-se as curvas obtidas. Quanto mais próxima do topo do gráfico a curva estiver, melhor será o resultado da operação de busca. Sendo assim, a curva ideal para uma consulta apresenta 100% de precisão para todos os valores de revocação.

3.5 Considerações Finais

Neste capítulo foram discutidos os principais tópicos relacionados aos sistemas médicos PACS, CAD e CBIR. Particularmente, os sistemas CAD e CBIR são explorados nesta tese, sendo que seus maiores desafios são: representar adequadamente as imagens através de vetores de características; determinar as características mais importantes para categorizar as imagens; determinar a melhor função de distância a ser utilizada para comparar as imagens; determinar uma técnica de mineração confiável e robusta para a descoberta de padrões em imagens. O último tópico é o centro de pesquisa desta tese e é discutido no próximo capítulo.

Capítulo 4

Mineração de Dados e de Imagens

4.1 Considerações Iniciais

Mineração de dados (*data mining*) é o processo de explorar grandes quantidades de dados à procura de informações e padrões ocultos. É um tópico da ciência da computação que combina esforços de várias áreas, como estatística, banco de dados, inteligência artificial e reconhecimento de padrões. A informação extraída através da mineração de dados pode ser usada em uma grande variedade de aplicações, tais como a análise de mercado, o controle de produção e a análise de dados biomédicos. Uma variante da mineração de dados, que explora padrões especificamente em imagens é a *mineração de imagens*.

A *mineração de imagens* usa os princípios da mineração de dados, porém de uma maneira mais complexa, envolvendo um conjunto de técnicas de processamento de imagens, extração de características, indexação de objetos complexos e visualização. Muitas pessoas confundem os termos *mineração de imagens* e *mineração visual*, no entanto esses termos têm significados distintos. A *mineração de imagens* refere-se à extração automática ou semi-automática de padrões de imagens. Já a *mineração visual* consiste em visualizar de uma maneira inteligente grandes conjuntos de dados com o intuito de identificar padrões nos mesmos.

Nesta tese, técnicas de mineração são usadas para a análise de imagens médicas. Primeiro, as técnicas de mineração são usadas para tornar a busca por conteúdo mais eficiente e mais precisa. Depois, as técnicas de mineração de imagens são utilizadas para a construção de um método de auxílio ao diagnóstico de imagens médicas. Assim, tem-se que a mineração constituí um dos pilares desta tese. A principal tarefa de mineração explorada neste trabalho foi a mineração de regras de associação. No entanto, também foram exploradas outras tarefas de mineração, como a classificação, a seleção de características, a discretização (pré-processamento dos dados) e a teoria fractal. Todos esses conceitos são discutidos neste capítulo.

Este capítulo está organizado da seguinte forma. Na seção 4.2, são apresentados o processo

de descoberta de conhecimento e mineração de dados. Na seção 4.3, são apresentadas as etapas do processo de mineração de imagens. A etapa de pré-processamento de dados, que inclui as técnicas de discretização e seleção de características, é discutida na seção 4.4. A tarefa de associação é discutida na seção 4.5. A tarefa de classificação é discutida na seção 4.6. A teoria fractal é brevemente apresentada na seção 4.7.

4.2 O Processo de KDD

A facilidade com que os sistemas computacionais geram uma imensa quantidade de dados fez com que se tornasse necessário o desenvolvimento de técnicas automáticas para extrair padrões sobre os dados armazenados. Assim, surgiu o processo de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD).

A definição mais usada de KDD foi fornecida por Fayyad [Fayyad et al., 1996]: "A descoberta de conhecimento é um processo não trivial de identificação de padrões que sejam válidos, novos, potencialmente úteis e compreensíveis". O processo de KDD refere-se às etapas que produzem conhecimento a partir dos dados e, principalmente, à etapa de mineração dos dados, que é a fase que os transforma em informações. Esse processo envolve encontrar e interpretar padrões nos dados, de modo iterativo e interativo, através da execução de algoritmos e da análise de seus resultados.

Fayyad [Fayyad et al., 1996] divide o processo de KDD em 9 fases, incluindo: conhecimento do domínio da aplicação, obtenção do conjunto de dados, limpeza e pré-processamento dos dados, redução do volume de dados, escolha da tarefa de mineração, escolha do algoritmo de mineração, mineração de dados, interpretação e avaliação dos padrões descobertos, e, utilização do conhecimento obtido. O processo de KDD pode ser compactado nas etapas apresentadas na tabela 4.1 e ilustradas na figura 4.1.

Tabela 4.1: Principais etapas do processo de KDD.

Etapa	Objetivo
Seleção	Buscar na base os dados relevantes para a tarefa de análise.
Pré-processamento	Eliminação de ruídos e erros e compatibilização dos dados.
Mineração dos dados	Aplicação de algoritmos para a extração de conhecimento.
Apresentação	Apresentação do conhecimento usando técnicas adequadas.

O *pré-processamento* dos dados é responsável pela eliminação de ruídos e erros nos dados e pela conversão do formato dos dados, se necessário. Nessa fase, tarefas importantes como a seleção de características e a discretização dos dados são executadas. A *seleção de características* é um processo onde os atributos com maior poder de representação dos dados são mantidos e os demais são eliminados. A *discretização* é um processo que muda o domínio dos dados de contí-

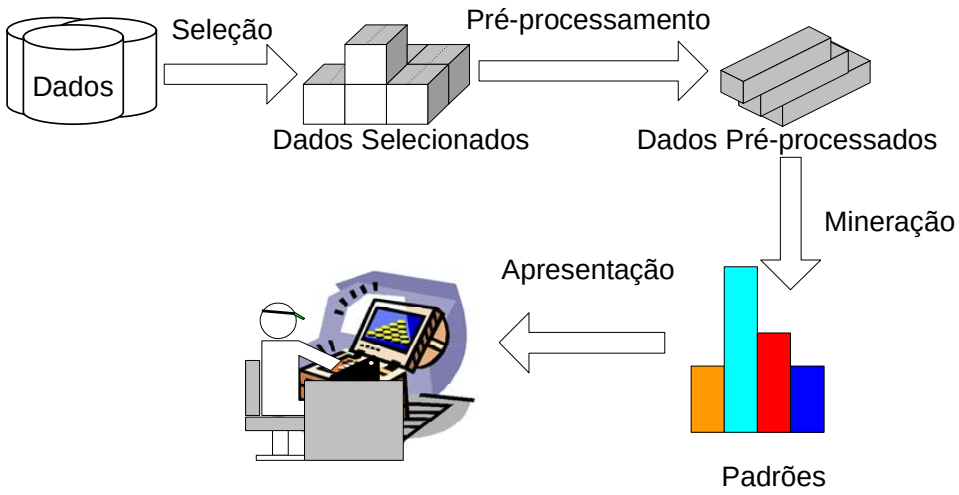


Figura 4.1: As etapas do processo de KDD.

nuo para discreto. Nesta tese, técnicas de discretização e seleção de atributos são desenvolvidas e utilizadas na busca por conteúdo e no método de auxílio ao diagnóstico desenvolvido.

A *mineração de dados*, que é considerada a parte principal do processo de KDD, tem sido intensamente discutida na literatura nos últimos anos. Ela constitui-se na etapa do processo de KDD onde efetivamente é feita a análise dos dados para a descoberta de padrões. Nessa fase, algoritmos são executados sobre os dados previamente preparados para a extração de conhecimento. Muitas vezes, devido à sua importância no processo, o termo mineração de dados é usado para se referir a todo o processo de KDD [Han & Kamber, 2000].

Uma **tarefa de mineração** é um conjunto de técnicas, procedimentos e algoritmos utilizados para a extração de um determinado tipo de conhecimento dos dados. O objetivo de minerar os dados é a verificação de uma hipótese ou a obtenção de novos padrões sobre os dados. Assim, as tarefas de mineração podem ser de *predição* ou *descrição*. Enquanto as tarefas de *predição* constroem modelos para predizer o comportamento de dados futuros, as tarefas de *descrição* revelam padrões e propriedades presentes nos dados analisados.

As principais tarefas de mineração são:

- Classificação: prediz a classe de um novo objeto;
- Agrupamento: agrupa objetos semelhantes;
- Associação: encontra relacionamento entre itens na base de dados;
- Sumarização: sintetiza os dados;
- Detecção de Desvios: procura por mudanças no comportamento dos dados.

Além das tarefas acima, outras tarefas de mineração estão surgindo com o progresso das pesquisas. As tarefas de mineração não são completamente disjuntas, muitas vezes, elas se inter-relacionam, onde técnicas desenvolvidas para uma tarefa podem ser aplicadas em outra e vice-versa. Um exemplo de inter-relacionamento entre as tarefas de mineração são os classificadores associativos, onde técnicas de associação são utilizadas para a classificação de novos dados [Thabtah, 2007].

Esta tese enfoca o emprego de técnicas de associação para o suporte da busca por conteúdo e para auxílio ao diagnóstico de imagens médicas. Além disso, um classificador associativo é desenvolvido para ser incorporado na última etapa do método desenvolvido de auxílio ao diagnóstico. Técnicas de classificação são utilizadas para comparar os resultados.

4.3 Mineração de Imagens

Um dos principais objetivos da análise de imagens pelo computador é tornar a máquina capaz de coletar informações relevantes automaticamente a partir das imagens, como a habilidade de extrair informações pertinentes a partir de um fundo com detalhes irrelevantes e a habilidade de apreender e fazer inferências a partir da informação incompleta. A mineração de imagens adiciona ao campo da mineração a complexidade de se trabalhar com imagens. Na mineração de imagens devem ser respondidas algumas questões que não se aplicam à mineração de dados tradicional, a saber:

- (a) Como extrair características visuais automaticamente das imagens?
- (b) Qual extrator usar?
- (c) É necessário um pré-processamento da imagem?
- (d) Qual será a representação da imagem?
- (e) Como ela será indexada?

A mineração de imagens tem sido foco de muitas pequisas atualmente. Um dos maiores desafios dos pesquisadores da área de mineração de imagens é como efetivamente relacionar características de baixo nível, automaticamente extraídas dos pixels das imagens, com a percepção do ser humano (alto nível). De acordo com [Hsu et al., 2002], as pesquisas no campo de mineração de imagens seguem duas direções:

- A direção de *domínio-específico*;
- A direção de *propósito geral*.

A direção de *domínio específico* foca técnicas de processamento de imagens, onde o objetivo principal é processar a imagem e extrair as características mais representativas para um tipo específico de imagens. A direção de *propósito geral* foca no desenvolvimento de algoritmos de mineração para reduzir o "gap semântico" não enfatizando nenhum tipo de imagem. Assim, as técnicas de propósito geral trabalham aumentando a precisão das técnicas de domínio específico, trabalhando de maneira complementar. Existem carências na identificação e representação das características relevantes para a análise das imagens e na criação de métodos, não só para extrair padrões significativos, mas também o fazê-lo de modo eficiente [Rangayyan, 2005].

De uma maneira geral, a mineração de imagens envolve um conjunto de quatro etapas apresentadas na figura 4.2 e descritas na tabela 4.2.

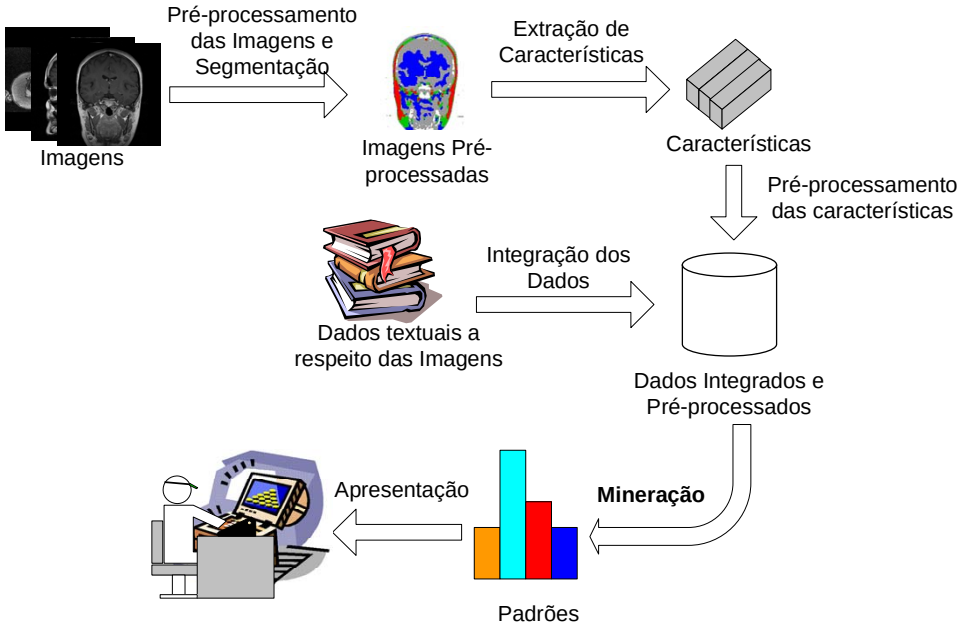


Figura 4.2: As etapas da mineração de imagens.

Tabela 4.2: As etapas da mineração de imagens.

Etapa	**Objetivo**
Pré-processamento	aumentar a qualidade da imagem e segmentá-la
Extração de características	extrair características automaticamente a partir da informação de pixels da imagem
Pré-processamento das características	eliminar ruídos e erros e compatibilizar o formato das características
Integração	integrar características visuais e dados textuais que descrevem as imagens para o processo de mineração
Mineração dos dados	aplicar algoritmos para extração de conhecimento
Apresentação	apresentar o conhecimento minerado usando técnicas adequadas

Além das características visuais extraídas dos pixels das imagens, também utiliza-se a descrição manual das imagens no processo de mineração, coletando o maior número possível de dados para a extração de padrões. No entanto, o fornecimento da descrição manual das imagens é um processo subjetivo, com alto grau de variabilidade sendo bastante lento e cansativo. No momento em que está fornecendo a descrição manual, o especialista pode não estar atento a alguns detalhes que, em outros momentos, foram descritos. Assim, embora a mineração de imagens utilize informações manualmente fornecidas a respeito das imagens, ela foca a utilização de técnicas automáticas para a extração de padrões. A fase de pré-processamento de imagens utiliza as técnicas de processamento de imagens e extração de características discutidas no capítulo 2.

4.4 Pré-processamento

Estudos recentes indicam que por volta de 80% do trabalho de descoberta de conhecimento se concentra na fase de pré-processamento [Zhang et al., 2005]. Assim, quanto mais bem preparado os dados, melhores são os resultados da mineração. Dessa forma, tem-se que o pré-processamento de dados é um passo crítico para o processo de descoberta de conhecimento. A preparação dos dados envolve limpeza, transformação, discretização e seleção. O principal objetivo do pré-processamento é providenciar dados de qualidade para a mineração de dados.

Freqüentemente, o pré-processamento se mistura com a própria mineração em si, os limites de onde termina um e onde inicia o outro nem sempre estão bem definidos. Por exemplo, a seleção de características é pré-processamento de dados ou mineração? Para promover seleção de características, algoritmos inteligentes são executados para identificar quais características são realmente relevantes e quais podem ser descartadas. Embora a seleção de características seja utilizada para melhorar o resultado de outros algoritmos de mineração, muitos autores a classificam como uma tarefa de mineração. A mesma questão pode ser feita sobre a discretização, pois ela freqüentemente requer a aplicação de algoritmos inteligentes sobre os dados.

4.4.1 Discretização

Os tipos mais comuns de atributos (características) usados na mineração de dados são os atributos *nominais* (categóricos), os *contínuos* e os *discretos*. Os atributos *nominais* freqüentemente assumem um número limitado de valores e não existe necessariamente uma relação de ordem entre os valores, um exemplo de atributo categórico é a cor do cabelo, por exemplo, preta, marrom ou vermelha. Já, os atributos contínuos podem assumir um infinito número de valores onde existe necessariamente uma relação de ordem entre eles. Um exemplo de atributo contínuo é o peso de uma pessoa. Os *atributos discretos* são atributos que possuem um número reduzido de

valores (em comparação com os atributos contínuos) e preservam a relação de ordem entre os valores. Ao mapeamento do domínio de um atributo contínuo para discreto, dá-se o nome de *discretização*.

A discretização de dados é um processo de dividir valores contínuos em intervalos. O objetivo de um algoritmo de discretização é determinar qual é o melhor conjunto de *pontos de corte* para discretizar os dados. Um ponto de corte é um limite de um intervalo de valores reais.

A discretização é uma etapa opcional que é executada no pré-processamento dos dados. Ela é utilizada para mapear valores contínuos em discretos quando é necessário reduzir o número de valores dos atributos e, ao mesmo tempo, manter a relação de ordem entre os valores. Muitas vezes, a discretização é necessária por exigência do algoritmo de mineração. Muitos algoritmos de mineração focam dados nominais, embora muitas aplicações reais necessitem lidar com dados contínuos, o que faz com que o uso de um algoritmo de discretização se faça necessário. Além dessas razões, a discretização também é utilizada para lidar com a distribuição espacial dispersa de valores e com o problema de se ter muitos valores distintos e poucos exemplos (instâncias de dados). Como desvantagem de utilizar discretização tem-se a perda de precisão dos valores e a perda de informação que ocorre durante o processo de discretização, que podem levar o algoritmo de mineração a ter resultados tortuosos. Apesar da perda de informação inerente ao processo de discretização, muitos trabalhos relatam um aumento significativo na precisão e na velocidade dos algoritmos de mineração ao utilizar uma técnica de discretização adequada na etapa de pré-processamento [Abraham et al., 2006] [Liu et al., 2002] [Kurgan & Cios, 2004].

As técnicas de discretização têm sido intensamente pesquisadas e diferentes tipos de discretização têm sido desenvolvidos. Alguns trabalhos [Liu et al., 2002] [Dougherty et al., 1995] foram desenvolvidos com o intuito de corretamente categorizar os métodos de discretização. Em geral, os métodos de discretização podem ser *supervisionados*, quando utilizam a informação da classe das instâncias (exemplos) para promover a discretização, e *não supervisionados*, quando não utilizam essa informação.

Os métodos que utilizam a discretização supervisionada utilizam o conceito de **classe majoritária** de um intervalo. A classe majoritária de um intervalo é a classe mais freqüente nas instâncias pertencentes ao intervalo. Outro conceito bastante usado é o de **inconsistência**. Uma *inconsistência* ocorre quando uma instância de uma classe diferente da classe majoritária é alocada em um intervalo.

Os métodos mais simples de discretização são o método do *tamanho fixo* (*equal-width*, intervalos de valores de mesmo tamanho) e o método da *freqüência fixa* (*equal-frequency*, intervalos com o mesmo número de instâncias), que não são supervisionados. Os métodos de *tamanho fixo* e de *freqüência fixa* não são indicados para serem aplicados para processar valores contínuos de distribuição não uniforme. Além disso, os métodos não supervisionados descartam a informação de classe das instâncias e, por isso, tendem a atribuir valores fortemente correlacionados

com a mesma classe para diferente intervalos, o que pode distorcer o resultado do algoritmo de mineração.

Um aprimoramento do método de *tamanho fixo* é o discretizador 1R proposto por Holte [Holte, 1993]. No método 1R, os limites dos intervalos (pontos de corte) são ajustados de acordo com a informação de classe das instâncias. Primeiro os valores contínuos são ordenados. Um parâmetro de entrada estabelece a freqüência mínima (número de instâncias) que um intervalo, exceto o último, deve possuir. Os ajustes dos limites são realizados utilizando a seguinte restrição: *um intervalo não pode ser dividido se a classe da próxima instância for igual a classe majoritária do intervalo*. O algoritmo 1R usa essa restrição para evitar a adição de inconsistência nos dados. Entretanto, este método de discretização produz um grande número de intervalos que pode ainda ser reduzido, conforme discutiremos no capítulo 6, onde é apresentado o algoritmo Omega desenvolvido durante o trabalho desta tese.

Em [Kerber, 1992], um algoritmo chamado ChiMerge foi proposto. O ChiMerge usa o teste estatístico χ^2 para determinar quando intervalos consecutivos devem ser agrupados. Um parâmetro estabelece o máximo valor que o teste χ^2 pode apresentar para unir (fundir) intervalos consecutivos. O algoritmo ordena os valores contínuos e, inicialmente, coloca cada valor separadamente em um intervalo. O valor de χ^2 é calculado separadamente para cada intervalo adjacente. Os intervalos adjacentes com o menor valor de χ^2 são unidos. Esse processo continua até que nenhuma fusão mais seja possível, ou até que um número mínimo pré-estabelecido de intervalos seja atingido. Experimentos apresentados em [Liu et al., 2002] comparando o ChiMerge com o discretizador 1R no pré-processamento para a tarefa de classificação, mostram que o 1R leva a uma maior precisão do classificador, além de permitir a construção de um modelo de classificação mais compacto. Além disso, o ChiMerge é mais custoso computacionalmente do que o 1R, pois o teste χ^2 é calculado para cada dois intervalos consecutivos para determinar se ocorre ou não a união dos mesmos, enquanto que para determinar a união de dois intervalos consecutivos, no 1R só é verificado se a classe da instância inicial de um intervalo é igual à classe majoritária do intervalo anterior.

Em [Fayyad & Irani, 1993] foi proposto um método que usa entropia e o princípio do comprimento mínimo da descrição (*Minimum Description Length Principle* - MDLP) para descartar pontos de corte não úteis. Pelo MDLP, o melhor caso de discretização é aquele que produz a maior compressão dos dados.

Em [Cerquides & Mantaras, 1997] foi proposto um método de discretização que realiza busca gulosa e usa uma distância específica, baseada na medida de entropia, para determinar os pontos de corte. Um algoritmo realiza uma *busca gulosa* quando ele escolhe sempre a opção local ótima em cada estágio com o intuito de encontrar a solução global ótima. No método proposto por [Cerquides & Mantaras, 1997], MDLP é o critério de parada na divisão de intervalos.

Em [Abraham et al., 2006], um estudo comparando a precisão do classificador Naive Bayes (discutido na seção 4.6), ao analisar dados médicos discretizados, concluiu que a discretização MDLP leva maiores valores de precisão do que os métodos de discretização por tamanho fixo e freqüência fixa.

Em [Liu & Wang, 2005] é apresentado um método que usa uma medida de heterogeneidade de classes. O método visa reduzir o número de intervalos enquanto maximiza o ganho de informação, e tem desempenho semelhante a outros métodos de discretização baseados em entropia.

Em [Ho & Scott, 1997] foi proposta uma medida chamada Zeta para determinar os pontos de corte no processo de discretização. De uma maneira simples, Zeta é a soma do número de ocorrências de instâncias da classe majoritária dos possíveis intervalos. Os pontos de corte escolhidos são aqueles que levam a um maior valor de Zeta e também que não separam instâncias consecutivas da mesma classe. O algoritmo termina quando um determinado número de intervalos é alcançado. Experimentos apresentados em [Liu et al., 2002] comparando Zeta com os discretizadores 1R, ChiMerge e MDLP no pré-processamento para a tarefa de classificação, mostram que Zeta levou à menor precisão do classificador.

Em [Liu & Setiono, 1997], um algoritmo chamado Chi2 que realiza seleção de características e discretização foi proposto. O Chi2 é uma versão melhorada do ChiMerge que usa o teste estatístico χ^2 para agrupar intervalos consecutivos. O Chi2 une intervalos consecutivos que levam ao menor valor de χ^2 em cada passo. Um critério de máxima *incoerência* é utilizado para determinar quando o algoritmo deve parar de unir intervalos. Duas instâncias são consideradas *incoerentes* se elas têm o mesmo valor de seus atributos, exceto o valor da classe. O uso do critério de *incoerência* permite uma tolerância a ruídos que leva a remoção de características irrelevantes. O processo de seleção de características é feito removendo o conjunto de características que geram apenas um intervalo; essas características tendem a ser independente de classes. O Chi2 tem complexidade algorítmica $O(N \log N)$ para discretizar um faixa de N valores ordenados. Nesta tese, foi desenvolvido um algoritmo chamado Omega (ver capítulo 6) que também realiza simultaneamente discretização e seleção de características. Como o Chi2 também realiza essas duas tarefas, a comparação experimental desses algoritmos é apresentada no capítulo 6.

Não foi encontrado na literatura trabalhos que aplicaram técnicas de discretização para a mineração de imagens médicas. Os artigos encontrados na literatura sobre esse assunto são relacionados ao trabalho de pesquisa desta tese [Ribeiro et al., 2007] [Ribeiro et al., 2008c] [Ribeiro et al., 2008b] e serão detalhados nos capítulos 6 e 7.

4.4.2 Seleção de Características

A *redução de dimensionalidade* é o processo de diminuir o número de características (atributos) usadas para representar o conjunto de dados em consideração. A redução de dimensionalidade visa diminuir o tamanho do vetor de características que representa os dados, por meio da remoção de características redundantes, correlacionadas e ruidosas. Na maioria dos casos, as técnicas de redução de dimensionalidade aumentam a velocidade e a precisão dos algoritmos de mineração de dados.

As técnicas de redução de dimensionalidade podem ser *supervisionadas* e *não supervisionadas*. As técnicas de redução de dimensionalidade também podem ser classificadas em técnicas de *seleção de características* e técnicas de *transformação de características*. Não existe consenso na literatura quanto a nomenclatura. De fato, as técnicas de *transformação de características* também são chamadas de técnicas de *extração de características* ou *técnicas de redução de características*.

A principal diferença entre as técnicas de *seleção* e *transformação* de características é que a primeira consiste em selecionar um subconjunto de características originais para representar os dados, enquanto a última consiste em gerar um conjunto reduzido completamente novo de características para representar os dados.

Nas técnicas de *transformação de características*, os valores das características são alterados para providenciar uma representação mais compacta. A *transformação de características* é obtida mapeando o espaço multidimensional para um espaço com um menor número de dimensões. As técnicas mais comuns de transformação de características emprega a Análise de Componentes Principais (*Principal Components Analysis, PCA*). A técnica PCA visa encontrar os eixos nos quais os dados apresentam a maior variância. Para isso, uma matriz de correlação dos dados é construída e os componentes principais (*eigenvalues*) são usados para transformar os dados originais, maximizando sua variância.

Nesta tese são enfocadas as técnicas de *seleção de características*. A *seleção de características* não transforma as características originais; ela somente remove o subconjunto de características redundantes e irrelevantes do conjunto de características. Sua grande vantagem é a preservação do significado semântico original dos dados. Foi mostrado em [Cheng et al., 2006] que as técnicas de seleção de características aumentam a especificidade de um classificador médico.

As técnicas de seleção de características podem ser divididas em *binárias* e *contínuas*. As técnicas contínuas de seleção de características atribuem pesos *contínuos* para cada característica, enquanto as técnicas *binárias* atribuem pesos binários para cada característica. As técnicas de seleção de características podem ainda seguir o modelo *filtro* ou o modelo *wrapper*. No modelo *filtro*, a seleção de características é realizada antes da fase de aprendizado e funciona

como um pré-processamento para o algoritmo de mineração. No modelo *wrapper*, o algoritmo de seleção de características usa o algoritmo de mineração como uma subrotina. A maior desvantagem do modelo *wrapper* é o alto custo computacional empregado pelo algoritmo de aprendizado para avaliar cada subconjunto de características [Molina et al., 2002]. De acordo com o método de busca empregado no algoritmo de seleção de características, o subconjunto de características para ser avaliado pode crescer, em métodos de *busca para frente* (*forward search*); ou diminuir, em métodos de *busca para trás* (*backward search*) até que um critério de parada seja atingido.

Um grande número de algoritmos de seleção de características tem sido apresentado na literatura. Um dos primeiros algoritmos foi apresentado em [Narendra & Fukunaga, 1977], onde a propriedade da *monotonicidade* (todos os subconjuntos de características relevantes são também relevantes) é empregada para podar o espaço de busca e uma medida de divergência é utilizada para avaliar o conjunto de características.

Um dos algoritmos de seleção de características mais conhecidos é o Relief [Kira & Rendell, 1992]. O princípio geral do Relief é medir a qualidade das características de acordo com a qualidade que os seus valores distinguem instâncias de diferentes classes. Dada uma instância S escolhida randomicamente de uma base de dados R, com k características (atributos), o Relief procura pelo vizinho mais próximo da mesma classe, que é chamado *nearest hit* H, e pelo vizinho mais próximo de classe diferente, chamado de *nearest miss* M. Ele atualiza o estimador de qualidade $W[F_i]$ de todas as características dependendo da diferença entre os valores das instâncias S, H e M. Esse processo é repetido n vezes, onde n é um parâmetro especificado pelo usuário. A complexidade de tempo do Relief é $O(nkN)$, onde N é o número de instâncias da base e k é o número de características. Relief retorna uma lista de características ordenadas de acordo com suas relevâncias, no entanto, ele não fornece uma indicação do número de características que deve ser removido. Uma limitação do algoritmo Relief é que ele trabalha somente com conjuntos de dados cuja classificação é binária. Essa limitação foi solucionada com o desenvolvimento do algoritmo Relief-F [Kononenko, 1994], que é uma extensão do algoritmo Relief para trabalhar com dados cuja classificação pode assumir múltiplos valores.

Outra técnica bem conhecida de seleção de características é o método baseado em árvore de decisão, ou *Decision Tree Method* (DTM) [Cardie, 1993]. A DTM adota uma busca para frente (*forward search*) para gerar subconjuntos de características, usando o critério de entropia para avaliá-los. A DTM executa o algoritmo C4.5 [Quinlan, 1993] que constrói uma árvore de decisão. Como uma árvore de decisão é uma seqüência de atributos que define o estado de uma instância, DTM seleciona as características que aparecem na árvore de decisão, como o melhor subconjunto de características.

Em [Huang & Dai, 2003], a distribuição qui-quadrado é usada para inferir a distribuição dos dados e para promover seleção de características. Entretanto, outros testes estatísticos também

podem ser usados para inferir a distribuição dos dados e para promover seleção de características. Nesta tese, o teste estatístico Z é utilizado para gerar regras de associação.

Uma comparação entre o desempenho dos métodos de seleção de características pode ser encontrada em [Refaeilzadeh et al., 2007].

Seleção de Características em Mineração de Imagens

Em [Dy et al., 2003] foi apresentado um método que utiliza seleção de características em duas etapas para a redução de dimensionalidade e o aumento da precisão nas buscas por conteúdo de imagens de pulmão. A primeira etapa da seleção de características é supervisionada e visa encontrar as características que melhor discriminam as imagens em classes. Isso é feito utilizando um algoritmo que é derivado do C4.5. Em uma segunda etapa é utilizado um método de seleção de características não supervisionado que realiza uma busca seqüencial, adicionando e avaliando uma característica por vez no subconjunto de características selecionadas. Para a seleção de características não supervisionada são usados agrupamentos nos dados encontrados usando o método EM (*Expectation Maximization*). Os resultados dessa técnica apresentados em [Dy et al., 2003] mostraram um ganho de precisão nas consultas de CBIR, no entanto, com um alto custo computacional.

Em [Poonguzhali et al., 2007], um método baseado em PCA foi utilizado para reduzir o vetor de características composto por características de textura para representar imagens de fígado. Após a redução de características, as imagens foram agrupadas utilizando o método *k-means*, onde os erros de agrupamento, baseado na informação das classes normal, cisto, benigno e maligno, foram reduzidos.

4.5 Associação

A tarefa de associação é uma tarefa que encontra relacionamentos entre a ocorrência de itens na base de dados. Por descrever o comportamento de dados já existentes, a tarefa de associação é considerada uma *tarefa de descrição*. Desde seu surgimento em 1993 [Agrawal et al., 1993], a associação se tornou uma técnica bastante utilizada por pesquisadores acadêmicos, devido a sua vasta aplicabilidade e a facilidade de compreensão dos padrões que ela gera. De fato, associações refletem como os seres humanos aprendem novos conhecimentos e por isso, a interpretação desse tipo de padrão é mais intuitiva.

O problema de minerar regras de associação foi introduzido em [Agrawal et al., 1993] como um problema de encontrar relacionamentos entre a ocorrência de itens em tuplas (registros) de uma tabela (base). Seja $I = \{i_1, \ldots, i_n\}$ um conjunto de literais, denominados itens. Um conjunto $X \in I$ é chamado de *itemset*. Um *itemset* X com k elementos é chamado de *itemset-k*. Seja R uma tabela com tuplas t que envolvem elementos que são subconjuntos de I. A tupla t

suporta um *itemset* X, se $X \in t$. Uma regra de associação é uma expressão da forma $X \rightarrow Y$, onde X e Y são *itemsets*. X é chamado de *corpo* ou antecedente da regra, e Y é chamado de *cabeça* ou conseqüente da regra. Seja $|R|$ o número de tuplas na tabela R. Seja $|Z|$ o número total de ocorrências do *itemset* Z nas tuplas da tabela R. As medidas de *suporte* sup (equação 4.1) e *confiança* $conf$ (equação 4.2) são utilizadas para minerar as regras de associação.

$$sup(X \rightarrow Y) = \frac{|X \cup Y|}{|R|} \tag{4.1}$$

$$conf(X \rightarrow Y) = \frac{|X \cup Y|}{|X|} \tag{4.2}$$

O problema de mineração de regras de associação, como ele foi inicialmente estabelecido envolve encontrar as regras que satisfazem as restrições de suporte mínimo (*minsup*) e confiança mínima (*minconf*) especificadas pelo usuário.

O suporte de um *itemset* X é a razão entre o número de tuplas em R que suportam X e o número total de tuplas de R. O suporte é utilizado como uma restrição da freqüência de *itemsets* para minerar as regras. Um *itemset* X é chamado *itemset freqüente* se o suporte de X for maior ou igual ao suporte mínimo especificado pelo usuário. Uma regra de associação $X \rightarrow Y$, onde $X \cap Y = \oslash$, pode ser traduzida como "se X então Y" indicando que quando X ocorre, Y tende também a ocorrer. A confiança de uma regra $X \rightarrow Y$ é a razão entre o número de tuplas que contém X e Y, e o número de tuplas que contém X. A confiança também é chamada de medida de "força" de uma regra.

Um exemplo bem conhecido [Agrawal et al., 1993]de regra de associação envolvendo dados de uma cesta de compras é "70% das compras que contêm fralda também contêm cerveja e 4% de todas as compras contêm esses dois itens". Nesse exemplo, 70% é a confiança da regra e 4% é o suporte da regra. As regras que satisfazem as restrições de suporte mínimo e de confiança mínima especificadas pelo usuário são chamadas de *regras fortes*.

O suporte possui a propriedade de *monoticidade*, que significa que todos os subconjuntos de itens de um conjunto freqüente também devem ser *freqüentes*. Assim, nenhum *itemset* obtido a partir de combinações de *itemsets* não-freqüentes pode ser *freqüente*. Essa propriedade é usada para podar o espaço de busca por *itemsets* freqüentes em alguns algoritmos de mineração, como o Apriori [Agrawal & Srikant, 1994]. A desvantagem do uso do suporte como medida de interesse é a eliminação de padrões com alta confiança e com baixa freqüência, que podem ser importantes em determinadas situações. Por exemplo, ao analisar dados de uma farmácia, pode-se ter que o medicamento A raramente é comprado, mas sempre que o mesmo é comprado, o medicamento B é vendido. Nesse caso, a restrição do suporte mínimo pode impedir que a regra $A \rightarrow B$, que possui um alto valor de confiança, seja encontrada.

O problema apresentado anteriormente é também conhecido como o problema de mineração

de regras de *associação booleanas*, assim denominado porque ele pode ser visto como um problema de encontrar associações entre valores "1" em uma tabela que tem um atributo booleano para cada item $i_i \in I$ e um registro r_j para cada tupla $t_j \in R$. Em um registro r_j, o valor "1" de um atributo indica que o item representado por esse atributo está presente na tupla t_j, e o valor "0" indica que o item está ausente.

As regras de associação geralmente são classificadas de acordo com as seguintes características:

- *Quanto ao tipo de atributo envolvido*: O domínio dos atributos de uma tabela submetida a um processo de mineração pode ser quantitativo ou categórico. Atributos quantitativos são atributos numéricos contínuos, onde existe uma ordem explícita entre seus valores, como idade e preço. Atributos categóricos são atributos que têm um domínio discreto e finito de elementos, não apresentando uma ordem explícita entre os valores dos elementos, como cor. Outro tipo de atributo usado nas regras de associação são os atributos nebulosos. Um atributo nebuloso é aquele cujo valor é um termo lingüístico impreciso ou incerto, como velho, novo e bastante. Assim, as regras de associação podem ser classificadas quanto ao tipo de atributo envolvido em: regras de associação booleanas (quando os atributos envolvidos são categóricos), regras de associação quantitativas (quando os atributos envolvidos são atributos numéricos contínuos) e regras de associação nebulosas (regras que envolvem conceitos nebulosos).

- *Quanto ao nível de detalhamento envolvido*: Os itens da base de dados podem ser agrupados em diferentes níveis de abstração. As regras de associação podem ser classificadas de acordo com o nível de abstração de seus itens em *regras de associação multiníveis* (quando envolvem itens em diferentes níveis de abstração) e regras de associação *uniníveis* (quando envolvem itens de apenas um nível de abstração).

- *Quanto ao relacionamento entre os itens da regra*: As regras de associação também podem ser classificadas quanto ao tipo de relacionamento existente entre seus itens. Uma regra de associação $X \rightarrow Y$, onde X e Y são conjuntos de itens, pode ser: *direta*: a presença de X aumenta a possibilidade da presença de Y em uma tupla; ou, *inversa ou negativa*: a presença de X diminui a possibilidade da presença de Y em uma tupla.

É possível saber se a relação entre duas variáveis é *direta* ou *inversa* por meio do cálculo do coeficiente de correlação. O coeficiente de correlação entre duas variáveis X e Y em um conjunto de dados com n valores (x_i,y_i) é calculado pela equação 4.3, onde $\bar{x}$ e $\bar{y}$ é a média aritmética dos valores de X e Y respectivamente.

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{onde} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i) \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}(y_i) \tag{4.3}$$

O coeficiente de correlação é um valor entre -1 e 1. Se o valor da correlação é zero, não existe dependência linear entre as variáveis. Valores próximos de 1 indicam *associação direta* entre as variáveis X e Y, enquanto valores próximos de -1 indicam *associação inversa*.

4.5.1 Algoritmos de Mineração de Regras de Associação

O problema de minerar regras de associação consiste em encontrar todas as regras de *associação fortes* (regras que satisfazem o suporte mínimo e a confiança mínima) em uma base de dados.

A restrição de minerar regras que satisfazem os valores estabelecidos de suporte mínimo e confiança mínima permite dividir o problema de mineração em duas etapas [Agrawal & Srikant, 1994]:

- Encontrar $\{L = X \subseteq I | X$ é freqüente$\}$, onde I é o conjunto de todos os itens da base de dados analisada. L é o conjunto de todos os *itemsets* freqüentes, juntamente com seus respectivos valores de suporte.

- Para todos os *itemsets* freqüentes $X \in L$, calcular a confiança de todas as regras $Y \rightarrow X - Y$, onde $Y \in X$, sendo que $Y \neq \oslash$, e eliminar todas aquelas que não satisfazem a confiança mínima.

A fase crítica da mineração é a fase de determinação dos *itemsets freqüentes*. Segundo Hipp, Güntzer e Nakhaeizadeh [Hipp et al., 2000], o problema de encontrar regras de associação pode ser reduzido a encontrar todos os *itemsets* freqüentes e seus respectivos valores de suporte, uma vez que, tendo encontrado os *itemsets* freqüentes, para determinar as regras, basta gerar as combinações internas de cada *itemset* freqüente e calcular a confiança de cada combinação, descartando aquelas combinações que não satisfazem a confiança mínima estabelecida. Em geral, a fase de geração das regras a partir do conjunto de *itemsets* freqüentes é comum para a maioria dos algoritmos.

Os primeiros algoritmos para determinação de itemsets freqüentes foram AIS [Agrawal et al., 1993] e o SETM [Houtsma & Swami, 1993]. Em [Agrawal & Srikant, 1994] foi apresentado o algoritmo Apriori, que, devido a sua simplicidade, hoje é o algoritmo mais conhecido e utilizado de mineração de regras de associação.

Algoritmo Apriori

O algoritmo Apriori [Agrawal & Srikant, 1994], descrito no Algoritmo 1, usa a propriedade de monotonicidade para realizar as podas. No algoritmo 1, L_k é o conjunto de *itemsets freqüentes* de tamanho k (os *itemsets-k* que satisfazem a restrição de suporte mínimo *minsup*). C_k é o conjunto de *itemsets candidatos* de tamanho k (os *itemsets-k* potencialmente freqüentes).

Algoritmo 1: Algoritmo Apriori.

Dados: Tabela com tuplas t, suporte mínimo *minsup*

Resultado: Conjunto de *itemsets* freqüentes

1 $L_1 = \{\text{itens freqüentes}\}$;
2 **para** $(k = 1; L_k \neq \emptyset; k++)$ **faça**
3 C_{k+1} = novos candidatos gerados a partir de L_k;
4 **para** *cada tupla t na base de dados* **faça**
5 Incremente o contador de todos os candidatos em C_{k+1} que estão contidos em t.
6 **fim**
7 L_{k+1} = candidatos em C_{k+1} que satisfazem *minsup*
8 **fim**
9 **retorna** $\cup_k L_k$

Na linha 1, o algoritmo conta o número de ocorrências de cada item e determina L_1 (conjunto de *itemsets-1* freqüentes). As linhas 2 a 8 consistem em determinar L_k (conjunto de *itemsets-k* freqüentes). Sempre, o conjunto L_k, é usado para gerar C_{k+1} (o conjunto de *itemsets-k* candidatos). Na linha 3 é feita a geração dos *itemsets* candidatos C_{k+1}. Para isso, é feita uma junção de L_k consigo mesmo, sendo que a condição de junção é que os $k-1$ itens dos dados de junção sejam os mesmos. Após a junção, o algoritmo Apriori faz uma verificação para cada *itemset* gerado, se ele possui subconjuntos de itens não freqüentes. Caso possua, o itemset é eliminado do conjunto de *itemsets* candidatos, caso contrário ele é adicionado a C_{k+1}. Nas linhas 4 a 6 é feita a contagem de cada *itemset-k* candidato, onde seu contador é incrementado de 1 para cada tupla em que ele aparece. Por último, somente os *itemsets-k* candidatos que têm suporte maior ou igual ao suporte mínimo são adicionados a L_k e retornados.

A fase da mineração que exige maior processamento é a determinação dos *itemsets* freqüentes. Novos algoritmos foram desenvolvidos para tornar essa fase mais eficiente, dentre eles destacam-se os algoritmos Partition [Savarese et al., 1995], FP-Growth [Han et al., 2000] e Eclat [Zaki et al., 1997].

Um problema da mineração de regras de associação é o grande número de regras geradas. Em [Yamamoto et al., 2008] foram desenvolvidas técnicas de visualização de *itemsets* para a análise visual dos mesmos, que permitem ao usuário selecionar os *itemsets* que ele tem mais interesse que apareçam nas regras.

Outra questão bastante importante na mineração de regras de associação refere-se à *medida de interesse* a ser usada. Uma *medida de interesse* é uma medida do grau de importância de uma regra e define quais regras serão ou não retornadas pelo algoritmo de mineração. Além das medidas de interesse mais conhecidas - suporte e confiança - (já discutidas nesta seção), outras

medidas de interesse têm sido utilizadas na mineração de regras de associação. Algumas dessas medidas de interesse são apresentadas na tabela 4.3.

Tabela 4.3: Medidas de interesse usadas na mineração de regras de associaciação.

Medida	Cálculo	Significado
toda-confiança [Omiecinski, 2003]	$allconf(X) = \frac{sup(X)}{max(sup(x \in X))}$	Todas as regras geradas a partir de X têm confiança pelo menos $allconf(X)$.
conviction [Brin et al., 1997]	$convi(X \rightarrow Y) = \frac{1-sup(Y)}{1-conf(X \rightarrow Y)}$	Compara a probabilidade de X ocorrer sem o Y com sua freqüência de ocorrência.
lift [McNicholas et al., 2008]	$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{sup(X \rightarrow Y)}$	*Lift* mede o quão mais freqüente X e Y ocorrem juntos, em relação ao esperado se eles forem estatisticamente independentes.

Atualmente, tem-se notado também uma preocupação crescente com a privacidade dos dados. Diversas abordagens estão sendo desenvolvidas na literatura onde o foco é preservação da privacidade dos dados na mineração de regras de associação [Veloso et al., 2003] [Wang & Wu, 2005]. Outro foco de pesquisa é a mineração multi-relacional de regras de associação, onde regras são extraídas a partir de dados provenientes de múltiplas tabelas [Ribeiro & Vieira, 2004]. Em [Kazemzadeh & Sartipi, 2006], Kazemzadeh e Sartipi propõem uma maneira de fazer com que os resultados da mineração de dados fiquem disponíveis em um sistema clínico para auxiliar no processo de tomada de decisão e chamam a atenção para a importância de aplicar as técnicas de mineração de dados para auxílio ao diagnóstico.

A principal limitação para a aplicação de mineração de regras de associação em imagens é que os dados envolvidos freqüentemente são de domínio contínuo (quantitativo). A maioria das técnicas apresentadas na literatura somente é aplicada para dados de domínio categórico, sendo necessário, em muitos casos, discretizar valores contínuos. Muitas vezes a aplicação de técnicas para discretização de dados não é desejada, sendo necessário o desenvolvimento de algoritmos de regras de associação que trabalhem diretamente com dados contínuos. A mineração envolvendo dados contínuos pode ser feita por meio da mineração de regras de associação estatísticas, que é detalhada a seguir.

4.5.2 Regras de Associação Envolvendo Dados Contínuos

Em [Srikant & Agrawal, 1996] a definição de regras de associação foi estendida para incluir dados contínuos (quantitativos). A base para essa nova definição foi incluir dados quantitativos considerando intervalos de valores numéricos. Assim, um item de uma regra de associação pode

ser um valor categórico ou uma faixa de valores numéricos. Um exemplo de regra de acordo com a definição de [Srikant & Agrawal, 1996] é:

⟨ Idade: 30..39 ⟩ e ⟨ Estado_civil: casado⟩ → ⟨ número_carros: 2⟩

Srikant e Agrawal [Srikant & Agrawal, 1996] fornecem um algoritmo que encontra regras envolvendo atributos contínuos, que foram previamente discretizados. Primeiramente, os atributos contínuos são discretizados em um determinado número de intervalos. Para promover a discretização o algoritmo combina valores contínuos consecutivos até que os intervalos formados por esses valores satisfaçam a restrição de *suporte máximo*. O suporte máximo é uma medida de interesse adicional usada para limitar o máximo suporte de um item (intervalo) a ser conseguido com a discretização. A medida que os valores consecutivos são combinados para aumentar o suporte do item (intervalo) gerado, os valores de confiança das regras contendo esse novo item tendem a diminuir. Para minimizar esse problema de perda de confiança, o algoritmo proposto por [Srikant & Agrawal, 1996], após atingir o suporte máximo, o intervalo discretizado passa a ser divido até que pelo menos uma regra em que ele esteja presente (no corpo) satisfaça a *confiança mínima* e ainda satisfaça o *suporte mínimo*. Em adição, [Srikant & Agrawal, 1996] define um filtro de interesse para reduzir o problema de um grande número de regras similares. Existem alguns problemas com a abordagem proposta em [Srikant & Agrawal, 1996] que a inviabiliza de ser usada no problema de mineração tratado nesta tese:

- O processo de discretização utilizado é extremamente lento, envolve a contagem de medidas de interesse para várias configurações de intervalos. O número de varreduras para a contagem das medidas de interesse do algoritmo de mineração aumenta exponencialmente com o número de atributos contínuos presentes na base. Para o problema de minerar imagens médicas, que freqüentemente envolve vetores de características de alta dimensionalidade, a aplicação desse algoritmo é inviável;

- O número de regras geradas tende a crescer extremamente rápido com o aumento do número de atributos contínuos considerados, o que, considerando o problema de mineração envolvendo imagens médicas, levaria a uma explosão no número de regras.

Uma alternativa a abordagem de Srikant e Agrawal é o uso de regras de associação estatísticas, que são regras de associação encontradas com base na distribuição dos valores dos atributos quantitativos. As regras de associação estatísticas foram inicialmente definidas em [Aumann & Lindell, 1999] e são utilizadas nesta tese para a mineração de padrões em imagens médicas. Um exemplo de regra de associação estatística é:

sexo=feminino → *salário-hora: média=R$7.9*, (*salário-hora: média geral = R$9.02*)

Uma regra de associação indica um relacionamento interessante ou uma tendência em uma base de dados. As *regras de associação estatísticas* atendem esse propósito localizando subconjuntos de dados da base de dados que possuem um comportamento inesperado. Assim, uma regra de associação estatística em sua forma mais geral pode ser expressada por:

subconjunto da população $\rightarrow$ *comportamento interessante*

Tem-se que para o caso categórico o *comportamento* da base é descrito por uma lista de itens e sua probabilidade de ocorrência. Assim, um comportamento interessante para o caso categórico é a incidência, maior que a usual, de um grupo de itens em conjunto na base de dados. Estatisticamente, essa descrição é a distribuição de probabilidades de um conjunto de itens para uma dada população. Dessa forma, para um conjunto de valores quantitativos, a melhor descrição de seu comportamento é a sua distribuição. Aumman e Lindell [Aumann & Lindell, 1999] sugerem o uso da *média* e da *variância* para descrever o comportamento de um atributo quantitativo e, para garantir a descoberta de padrões interessantes definem que: "um subconjunto tem um comportamento interessante se sua distribuição é diferente do restante da população". Esse comportamento interessante pode ser descrito em termos de várias medidas que refletem a distribuição estatística, por exemplo, a média, a mediana e a variância. As regras de associação estatísticas foram inicialmente definidas como:

$X \rightarrow f_i = média_{T_x}(f_i)$

Seja T o conjunto de todas as transações da base de dados analisada. O *antecessor* (cauda) da regra X é um conjunto composto por valores de atributos categóricos e intervalos de atributos quantitativos discretizados. O *sucessor* (corpo) $f_i = média_{T_x}(f_i)$ é composto por um atributo quantitativo f_i e a média de f_i no conjunto $T_x \subseteq T$ de transações da base que satisfazem X. A regra $X \rightarrow f_i = média_{T_x}(f_i)$ é interessante se $média_{T_x}(f_i)$ é significantemente diferente da média $média_{T-T_x}(f_i)$. Embora os valores das médias em T_x e $T - T_x$ possam se diferenciar numericamente, é possível não haver nenhuma evidência estatística para inferir que existe realmente uma diferença real entre as populações. Assim, testes de hipóteses estatísticos podem ser usados para estabelecer o nível de significância da diferença entre as populações T_x e $T - T_x$. Exemplos de testes de hipóteses estatísticos que podem ser aplicados sobre os dados: o *teste-Z*, o *teste-F*, *teste-*χ^2 e o *teste-T*.

A definição de regras de associação estatísticas é uma generalização da definição de regras de associação booleana. Considere que $Z = X \cup Y$ seja um conjunto de itens. Seja A uma variável aleatória, onde $A = 0$, se Z não ocorre (fracasso); e $A = 1$, se Z ocorre em uma tupla da base de dados (sucesso). Seja p a probabilidade de sucesso e q a probabilidade de fracasso, onde $p + q = 1$. Esse comportamento indica que a ocorrência de um *itemset* assume distribuição de Bernoulli. Para a distribuição de Bernoulli, a função de probabilidade é dada por:

$$P(A = a) = p^a q^{1-a} \tag{4.4}$$

e tem-se que:

$$média(A) = E(A) = p \tag{4.5}$$

onde $E(A)$ é a *esperança* de A.

Nesse tipo de distribuição a média é dada por p (probabilidade de sucesso) que é a probabilidade de ocorrência do itemset que forma a regra $Z = X \cup Y$ na base de dados. Para uma regra de associação $X \rightarrow Y$, tem-se que p = suporte($X \cup Y$) e a confiança($X \rightarrow Y$) é a probabilidade de Y dado X, que é média Y no subconjunto da base onde ocorre X. Mapeando para o caso de regras quantitativas, o uso da média de um valor quantitativo como sucessor na regra de associação estatística, corresponde à medida de confiança da regra de associação booleana.

De uma maneira geral, o termo *regras de associação estatísticas* é usado para representar regras de associação que utilizam testes estatísticos para confirmar sua validade. Um teste estatístico é geralmente utilizado para determinar se existem evidências estatísticas de que uma hipótese feita sobre os dados pode ser rejeitada com um certo grau de confiança. Um exemplo que ilustra a importância de um teste estatístico para determinar se as médias são estatisticamente diferentes é apresentado na figura 4.3. A figura 4.3 apresenta três exemplos de distribuição de duas amostras (que seguem a distribuição Normal) em estudo. Note que, nos três exemplos a diferença entre as médias é a mesma. No entanto, pode-se concluir que as amostras da figura 4.3 (a) são as mais diferentes, pois existe uma pequena área de sobreposição entre as curvas. No caso da figura 4.3 (c) existe pouca diferença entre as amostras, pois as distribuições têm uma grande área de sobreposição. Apesar das diferenças entre as médias serem iguais, a população das amostras da figura 4.3 (a) é mais estatisticamente diferente em relação aos demais exemplos da figura, e essa diferença estatística pode ser determinada através da aplicação de um teste de hipóteses.

O *teste-Z* é um teste estatístico usado para inferir se a diferença entre as médias de uma amostra e de uma população, que seguem a distribuição Normal, é grande o suficiente para ser estatisticamente significante, de maneira a ser improvável que essa diferença tenha ocorrido por coincidência. Muitos dos fenômenos aleatórios de interesse se comportam próximos a distribuição Normal, com valores muito freqüentes em torno da média, diminuindo a freqüência à medida que se afastam da média. A distribuição Normal também é utilizada como aproximação de outras distribuições. Nesta tese, o *teste-Z* foi utilizado para a mineração de regras de associação estatísticas.

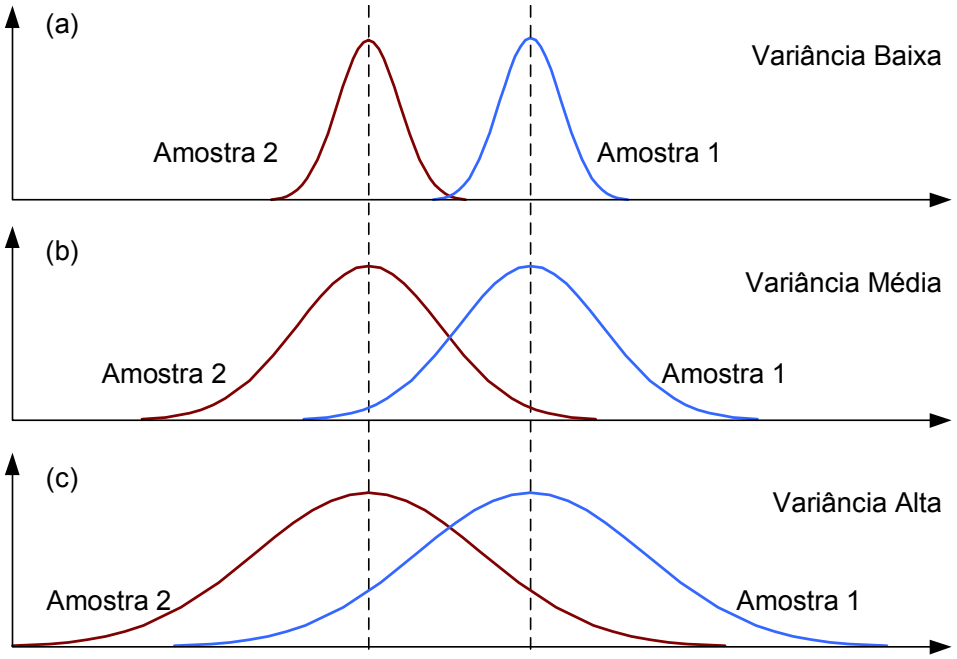


Figura 4.3: Três exemplos de distribuição de duas amostras em estudo.

4.5.3 Mineração de Regras de Associação em Imagens e para o Auxílio ao Diagnóstico

Um procedimento para a descoberta de regras de associação envolvendo o conteúdo de um conjunto de imagens com objetos simples foi proposto em [Ordonez & Omiecinski, 1999]. Primeiro, a imagem é segmentada em regiões chamadas de "*blobs*", usando o método EM (*Expectation Maximization*). Um vetor de características é gerado para representar cada *blob*. Uma função de distância é aplicada para comparar a similaridade entre os *blobs* de diferentes imagens. Os *blobs* considerados similares são representados pelo mesmo identificador de objetos (OID). O conjunto de OIDs referente a objetos que compõem uma determinada imagem são usados para formar o registro (tupla) que representará a imagem durante o processo de mineração. Um algoritmo de mineração de regras de associação é aplicado sobre os registros das imagens, gerando regras que relacionam os identificadores dos objetos. As regras resultantes mostram o relacionamento entre os objetos mais freqüentes. A maior restrição deste método é sua baixa aplicabilidade, pois somente pode ser aplicado em imagens que possuem formas simples. Assim, esse método não pode ser aplicado para a mineração de imagens médicas, que envolve um grande número de objetos complexos.

Ordonez et al. [Ordonez et al., 2006] definiu algumas restrições e regras de sumarização aplicadas a mineração de regras de associação envolvendo dados médicos. As regras de associação são filtradas usando suporte, confiança e *lift*, onde *lift* ajuda na seleção de regras com alto poder de predição. Dados numéricos (como idade e pressão) são discretizados e é o especialista

que determina o intervalo apropriado. Enquanto o método proposto por Ordonez, trabalha somente com informações de alto nível (fornecida manualmente) a respeito das imagens, o *método* desenvolvido nesta tese também trabalha com dados de baixo nível automaticamente extraídos das imagens.

Um *framework* para obter regras de associação relacionando objetos e categorias de tumores cerebrais foi proposto em [Pan et al., 2005]. O *framework* usa um método derivado da segmentação *watershed* guiado por um especialista no domínio para detectar as regiões de interesse (ROIs). Para cada ROI, a descrição textual de características de forma e tamanho, entre outras, são utilizadas para construir uma tabela com as transações das imagens. Cada tupla na tabela tem características de dois objetos, suas posições relativas, e um descritor da imagem indicando se os objetos são normais ou anormais. Restrições definidas por especialistas de domínio são usadas para restringir a ocorrência de alguns *itemsets* no corpo ou na cabeça das regras, reduzindo o número de regras geradas, que tende a ser bastante grande. O maior problema desse *framework* é que ele requer que o trabalho difícil e subjetivo de classificar manualmente os objetos e características das imagens.

Regras de associação são utilizadas em [Wang et al., 2004] para classificar mamografias. Primeiro, três características de forma são extraídas para cada imagem. Um registro combinando as características das imagens com sua classificação (benigna ou maligna) é gerada para cada imagem. As características são discretizadas em dez intervalos de tamanho fixo para serem posteriormente submetidas a um algoritmo de mineração de regras de associação. As regras são mineradas com a restrição de não apresentarem itens de classificação em seu corpo. Uma nova imagem é classificada de acordo com o número de regras e a confiança das regras que ela satisfaz. Uma desvantagem desta técnica é o processo de discretização, que utiliza intervalos de tamanho fixo, podendo resultar em uma alta perda de informação. Conforme é apresentado no capítulo 6, a aplicação de um método inadequado para a discretização dos dados pode piorar, e em muito, a eficácia do algoritmo de mineração.

Um classificador associativo foi apresentado em [Antonie et al., 2003] [Zaiane et al., 2002]. Na fase de pré-processamento as imagens são recortadas e suavizadas usando equalização de histograma. Características de média, variança, distorção (*skewness*) e descontinuidade (*kurtosis*) são extraídas das imagens, e junto com alguns outros descritores (como posição da mama e tipo de tecido), formam os registros das imagens que são submetidos ao algoritmo Apriori. As regras são mineradas utilizando valores baixos de confiança e a ocorrência de itens referentes a classe fica restrita à cabeça das regras mineradas. As regras opcionalmente podem ser generalizadas e podadas. Dada uma nova imagem, o classificador contabiliza o número de regras que ela satisfaz e classifica a nova imagem. A grande desvantagem deste método é o valor baixo de confiança permitido para a mineração das regras. O uso de uma baixa confiança pode gerar um grande número de regras distorcidas que podem induzir o classificador a cometer um grande

número de erros, além de tornar o processo de classificação bastante lento.

Olukunle and Ehikioya [Olukunle & Ehikioya, 2002] descrevem problemas relativos a complexidade de análise das imagens médicas e propõe um algoritmo para acelerar a fase de mineração de regras de associação. A maior desvantagem dessa técnica é que as imagens necessitam ser previamente associadas a descritores, e a maioria dos descritores necessitam do auxílio de um especialista para serem associados às imagens.

No capítulo 7 é apresentado o *método* de auxílio ao diagnóstico desenvolvido nesta tese, que lida com grande parte das limitações dos métodos discutidos nesta subseção.

4.6 Classificação

A tarefa de classificação permite agrupar dados em uma hierarquia de classes, de acordo com os valores dos seus atributos. Os registros agrupados em classes são formados por um atributo alvo (ou atributo de classe), que determina a classe do registro, e um conjunto de atributos de predição. O objetivo é descobrir as relações existentes entre os atributos de predição e o atributo alvo, utilizando registros cuja classificação é conhecida. A tarefa de classificação é uma tarefa de predição, uma vez que ela prediz os valores do atributo alvo.

Na figura 4.4 é ilustrado o processo de classificação. A mineração de regras de classificação é feita em duas etapas. Na primeira etapa um algoritmo de classificação é aplicado sobre uma amostra do banco de dados, que é chamada de conjunto de treino. O conjunto de treino contém registros do banco cuja classificação é conhecida, ou seja, apresenta dados com o valor do atributo alvo preenchido. O resultado da execução da primeira etapa é um conjunto de regras de classificação. Na segunda etapa, as regras de classificação mineradas anteriormente são utilizadas para classificar os demais registros da base de dados.

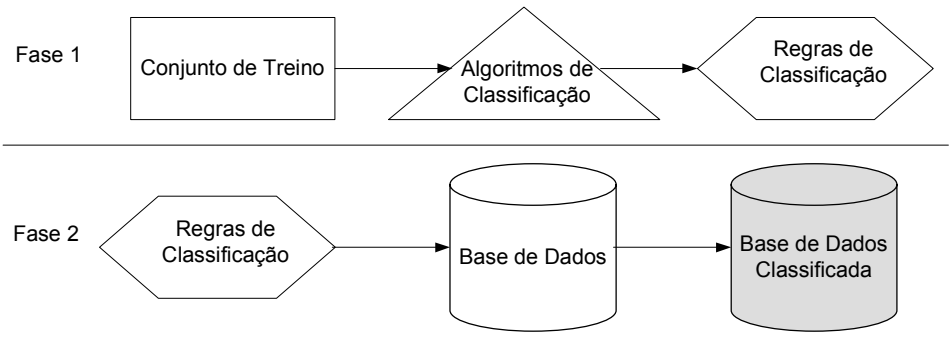


Figura 4.4: As duas fases do processo de classificação.

Um exemplo do processo de mineração de regras de classificação é fornecido a seguir. Considere os dados da tabela 4.4 como sendo um conjunto de treino com dados relativos ao sucesso

de lançamentos de novos aparelhos de barbear. A partir dos dados da tabela 4.4 é possível inferir as seguintes regras de classificação:

- Se (NROLÂMINAS = 2) e (PREÇO = barato) então SUCESSO = sim;
- Se (NROLÂMINAS = 1) então SUCESSO = não;
- Se (NROLÂMINAS = 3) e (PUB_ALVO = FEMININO) então SUCESSO = sim.

Tabela 4.4: Dados relativos ao sucesso de lançamentos de novos aparelhos de barbear.

NroLâminas	Pub_alvo	Preço	Sucesso
1	F	Caro	Não
2	F	Barato	Sim
3	F	Caro	Sim
3	M	Barato	Sim
2	M	Caro	Não
2	M	Barato	Sim
1	F	Barato	Não
2	F	Caro	Sim
3	F	Barato	Sim

Existe um conjunto de técnicas usadas para a obtenção de regras de classificação, sendo que as mais utilizadas são as árvores de decisão, as redes neurais e a classificação bayesiana. Essas técnicas são discutidas a seguir.

4.6.1 Árvores de Decisão

Amplamente utilizadas em algoritmos de classificação, as árvores de decisão são representações simples e eficientes do conhecimento. Uma árvore de decisão tem a função de particionar recursivamente um conjunto de treino, até que cada subconjunto obtido deste particionamento contenha casos de uma única classe [Han & Kamber, 2000]. Para atingir essa meta, a técnica de árvores de decisão examina e compara a distribuição de classes durante a construção da árvore.

A árvore de decisão é uma estrutura em forma de fluxograma, onde os nós internos representam um teste sobre o valor de um atributo e os nós folhas representam as classes. Um novo caso é classificado seguindo o caminho da raiz até as folhas. A figura 4.5 mostra uma árvore de decisão construída a partir da amostra de treino apresentada na tabela 4.4, onde os nós folhas indicam o valor do atributo sucesso para o lançamento de um novo aparelho de barbear.

Na construção de uma árvore de decisão, inicialmente todos os dados de treino estão na raiz. Dividem-se os dados de treino recursivamente através dos atributos selecionados até que seja satisfeita uma das seguintes condições de parada:

- Todos os dados de um mesmo nó pertencem a uma mesma classe;

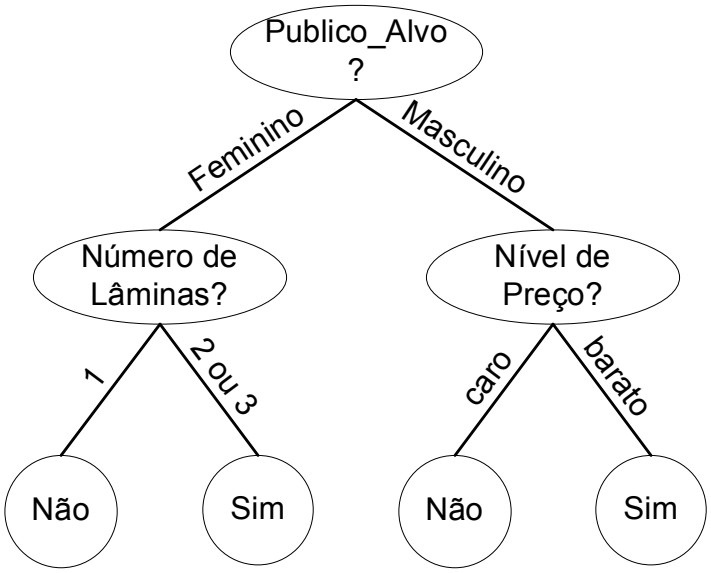


Figura 4.5: Exemplo de árvore de decisão.

- Não há mais atributos sobrando para o particionamento;

- Não há mais dados de treino.

Um dos principais problemas da construção de árvores de decisão é como selecionar o melhor atributo para fazer o teste para dividir as amostras (dados de treino). Um dos critérios mais usados para a construção de uma árvore de decisão é selecionar os atributos que resultam no maior ganho de informação. Quando um atributo é selecionado para ser colocado em uma árvore de decisão, as amostras são divididas de acordo com o valor desse atributo e o ganho de informação obtido com a escolha do mesmo é calculado da seguinte maneira:

ganho de informação = entropia antes da divisão - entropia depois da divisão,

onde a medida de entropia é a medida de desordem de um sistema, calculada de acordo com o apresentado a seguir. Seja S um conjunto de amostras com i classes. Seja $p_i = n_i/n$ a probabilidade de uma amostra ser da classe i, onde n_i é o número de amostras da classe i e n é o total de amostras. A entropia de um conjunto de dados é dada por:

$$entropia(p_1, p_2, ..., p_n) = -\sum_{i=1..n} p_i \log_2(p_i) \tag{4.6}$$

Um dos classificadores mais conhecidos baseado em árvores de decisão é o algoritmo C4.5 [Quinlan, 1993]. Esse classificador é utilizado nos experimentos desta tese para a validação dos métodos desenvolvidos.

4.6.2 Classificação Bayesiana

Um classificador bayesiano é um classificador estatístico baseado no teorema de Bayes. O classificador simples bayesiano (*Naive Bayes*) assume que o valor de um atributo é independente do valor dos demais atributos (independência condicional).

O teorema de Bayes é descrito a seguir. Seja X um exemplo cuja classificação é desconhecida. Seja H a hipótese de que X pertence à classe C. Para classificar o exemplo é necessário determinar $P(H \mid X)$, que é a probabilidade da hipótese H ser verdadeira, dado o exemplo observado X. Essa probabilidade é também chamada de probabilidade posterior ou probabilidade condicional, isto é, a probabilidade de H dado que X ocorreu. Similarmente, $P(X|H)$ é a probabilidade de X dado que H ocorreu. O teorema de Bayes é:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{4.7}$$

Considere que existam m classes C_1,C_2,...,C_m e que um exemplo de dado seja um vetor de características multidimensional, $X = (x_1, x_2, ..., x_n)$. Dado um exemplo cuja classificação é desconhecida, o classificador prediz que X pertence à classe que tem a maior probabilidade condicionada em X. Assim, o classificador atribui um exemplo X à classe C_i, se e somente se:

$$P(C_i|X) \geq P(C_j|X), 1 \leq j \leq m, j \neq i \tag{4.8}$$

Assim, deve-se encontrar a classe C_i, que tenha o valor $P(C_i|X)$ maximizado. Pelo teorema de Bayes tem-se:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{4.9}$$

Como $P(X)$ é constante para todas as classes, o valor de $P(X|C_i)P(C_i)$ deve ser maximizado, onde $P(C_i) = \frac{s_i}{s}$, sendo s_i o número de exemplos de treino da classe C_i e s o número total de exemplos de treino. O classificador *Naive Bayes* assume que não existe relação de dependência entre os valores dos atributos. Assim:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{4.10}$$

O valor de $P(x_k|C_i)$ pode ser estimado dos exemplos de treino da seguinte maneira:

- Se x_k for categórico, $\mathrm{P}(x_k \mid C_i) = \frac{s_{ik}}{s_i}$, onde s_{ik} é o número de exemplos de treino da classe C_i que teve o valor x_k para o atributo k e s_i é o número de exemplos de treino da classe C_i;
- Se o atributo k for contínuo, é assumido que ele possui uma distribuição *Gausiana* e então é calculada a probabilidade de acordo com a fórmula 4.11

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sigma_{C_i}\sqrt{2\pi}} e^{-\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \quad (4.11)$$

Na fórmula 4.11, $g(x_k, \mu_{C_i}, \sigma_{C_i})$ é a função de densidade *Gausiana* ou função de densidade normal para o atributo de índice k, enquanto μ_{C_i} e σ_{C_i} são respectivamente a média e o desvio padrão dos valores do atributo de índice k para os exemplos da classe C_i.

O classificador *Naive Bayes* é utilizado nos experimentos desta tese para a validação dos métodos desenvolvidos.

Além das técnicas de classificação discutidas nesta seção, as redes neurais e os algoritmos genéticos são também bastante empregados para a classificação. Uma rede neural é um conjunto de unidades de entradas e saídas conectadas, onde cada conexão tem um peso associado a ela. Durante a fase de aprendizado, a rede "aprende" ajustando os valores de seus pesos. Já os algoritmos genéticos se baseiam na teoria da evolução das espécies, envolvendo três operações: seleção, cruzamento e mutação. Usando essas três operações, os algoritmos genéticos continuamente atualizam a população corrente. Em cada iteração, cada população individual é avaliada de acordo com uma função que mede a aptidão do indivíduo no ambiente. O objetivo do algoritmo genético é evoluir os indivíduos mais aptos.

4.7 Fractais

A teoria dos fractais tem sido utilizada, dentre outras aplicações, na mineração de dados, para obter correlações entre características (atributos) [Sousa et al., 2002], para a seleção de características [Traina Jr. et al., 2000b], para a análise de *data streams* [Sousa et al., 2006] e para o reconhecimento de padrões [Costa & Jr., 2001].

Um fractal é definido pela propriedade de auto-similaridade, isto é, apresenta as mesmas características para diferentes variações em escala e tamanho. Assim sendo, partes do fractal - seja ele uma estrutura, um objeto ou um conjunto de dados - são similares, exatamente ou estatisticamente, ao fractal como um todo.

O termo fractal foi definido em 1975 por Benoit Mandelbrot, matemático francês que desenvolveu a **geometria fractal** na década de 1970. Um exemplo de fractal é ilustrado na Figura 4.6. A partir de um triângulo preenchido em um espaço 2D, um recorte triangular ao centro do mesmo determina a existência de três novos triângulos menores. Aplicando-se o mesmo recorte nos triângulos menores, em um processo recursivo de duração indefinida, o fractal conhecido como Triângulo de Sierpinski é criado. Os fractais podem ser conjuntos que possuem uma auto-similaridade perfeita, ou conjuntos que são aproximadamente auto-similares. De acordo com [Faloutsos & Kamel, 1994], em geral, dados reais podem ser modelados como fractais.

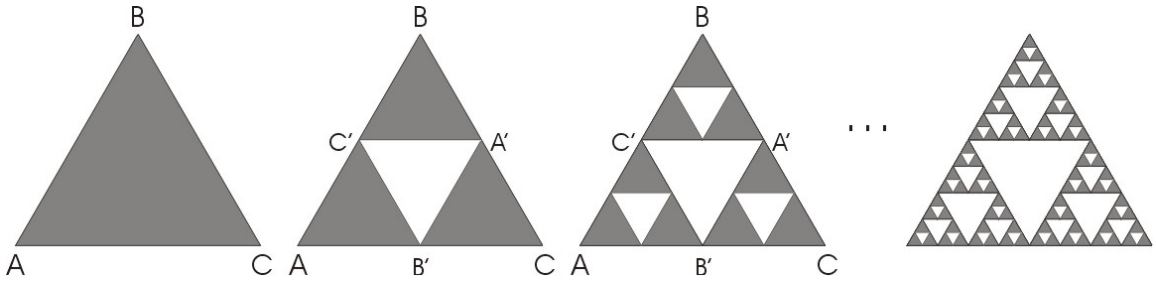


Figura 4.6: Exemplo de fractal: Triângulo de Sierpinski.

Na geometria dos fractais usa-se a medida de *dimensão fractal* para a interpretação estatística dos fractais, que indica o quanto um determinado fractal preenche o espaço no qual ele está imerso. O triângulo de Sierpinski da figura 4.6, por exemplo, é um fractal imerso no espaço 2D mas não é considerado um objeto de duas dimensões, pois possui área tendendo a zero. Por outro lado, ele também não é considerado um objeto unidimensional, pois possui perímetro tendendo ao infinito. Assim, é intuitivo pensar que o fractal está em algum lugar entre a linha e o plano, ou seja, ele possui uma dimensão fractal fracionária entre 1 e 2.

A dimensão fractal pode ser utilizada para revelar a dimensão intrínseca de um conjunto de pontos. A dimensão intrínseca representa a dimensão real do objeto e não a dimensão do espaço no qual o mesmo está imerso. Por exemplo, considere um conjunto de pontos, dispostos de maneira a formar um segmento de reta perfeito em um espaço tridimensional. A dimensão de imersão do conjunto é 3, mas sua dimensão intrínseca é 1. Ou seja, o conjunto pode ser perfeitamente transferido para o espaço uni-dimensional, sem que as distâncias entre os pontos sejam alteradas.

Na literatura, existem várias definições de dimensão fractal e várias maneiras de calculá-la. Em particular, as dimensões fractais de Hausdorff (D_0) e de correlação (D_2) têm sido utilizadas com sucesso em ferramentas de análise da distribuição dos dados, principalmente em trabalhos de pesquisa nas áreas de indexação e mineração de dados [Sousa, 2006].

Sendo M o número de réplicas e s o fator de escala segundo o qual cada réplica está reduzida, a *dimensão fractal* D de um fractal exatamente auto-similar definido em um espaço E-dimensional é:

$$D \equiv \frac{\log(M)}{\log s} \tag{4.12}$$

O Triângulo de Sierpinski, por exemplo, é um fractal exatamente auto-similar, pois sua regra de construção gera 3 réplicas em escala 1:2 a cada iteração. Logo, a dimensão fractal do Sierpinski é $D = \frac{\log(3)}{\log(2)} \approx 1.58$. Esta definição de dimensão fractal D é adequada para fractais matemáticos exatamente auto-similares, entretanto, para os conjuntos chamados fractais esta-

tisticamente auto-similares, é mais eficiente calcular a dimensão fractal de Hausdorff $D0$ utilizando o método Box-Counting [Traina Jr. et al., 2005]. Seja P um conjunto de pontos imerso em um espaço E-dimensional dividido por um hiper-reticulado com células de lado r, e seja $N(r)$ o número de células que contêm um ou mais pontos do conjunto. A *dimensão fractal de Hausdorff*, é definida por:

$$D0 \equiv -\lim_{r \to 0} \frac{\log(N(r))}{\log(r)} \tag{4.13}$$

Intuitivamente, a dimensão fractal de Hausdorff mostra como o número de "buracos" no fractal cresce conforme a granularidade se torna mais fina [Korn et al., 2001].

A dimensão de correlação D_2 é uma das formas mais simples de se computar a dimensão fractal. Dado um hiper-reticulado de células i de lado r, e seja $p_{i,r}$ a contagem de ocupação, isto é, a quantidade de pontos que incidem na i-ésima célula, a *dimensão fractal de correlação* D_2 é dada por:

$$D2 \equiv \frac{\partial \log(\sum_i (P_{i,r})^2)}{\partial \log(r)}, \; r \in (r_1, r_2) \tag{4.14}$$

onde r_1 e r_2 correspondem a mínima e a máxima distância entre dois pontos quaisquer.

No grupo GBDI a dimensão fractal tem sido calculada utilizando o algoritmo LiBOC, descrito em [Traina Jr. et al., 2005]. O algoritmo LiBOC permite o cálculo eficiente (custo linear no número de elementos do conjunto) da dimensão fractal de correlação D_2 e, conseqüentemente, fornece uma estimativa da dimensão intrínseca D de conjuntos de dados multidimensionais pontuais.

4.8 Considerações Finais

Nesta tese, as técnicas de mineração (de dados e de imagens) são usadas para tornar a busca por conteúdo mais eficiente e mais precisa e para a construção de um método de auxílio ao diagnóstico de imagens médicas. Assim, tem-se que a mineração constituí-se em um dos pilares desta tese. Neste capítulo foram discutidos os principais tópicos de mineração de dados e imagens que são explorados nesta tese. A principal tarefa de mineração explorada neste trabalho é a mineração de regras de associação. No entanto, também são explorados outras tarefas de mineração, como a classificação, a seleção de características, a discretização e a teoria fractal.

Nos capítulos anteriores foram vistos conceitos de processamento de imagens, de sistemas de suporte à área médica e neste capítulo foram discutidas as técnicas de mineração. Neste ponto já foram discutidos os tópicos relacionados com esse trabalho de pesquisa, assim os próximos capítulos discutem os trabalhos desenvolvidos e os resultados alcançados.

Parte II

Trabalhos Desenvolvidos

Capítulo 5

O Algoritmo StARMiner

5.1 Considerações Iniciais

As técnicas de CBIR usam vetores de características, no lugar das imagens, na execução das consultas por similaridade. Geralmente, em CBIR busca-se coletar o maior número de características visuais possíveis para representar as imagens, produzindo vetores de características com centenas ou até milhares de características. No entanto, a alta dimensionalidade dos vetores de características levam os sistemas de CBIR a sofrerem com o problema da "maldição da alta dimensionalidade": conforme o número de características aumenta, o poder de representação de cada característica diminui, e o processo de indexação e recuperação se degrada, tornando-se mais lento. Além disso, freqüentemente, um grande número de características são correlacionadas, trazendo informações redundantes e até ruídos que deterioram a habilidade do sistema distinguir corretamente as imagens. Assim, é importante manter o número de características o mais baixo possível, estabelecendo uma relação de compensação entre a capacidade de representação do vetor e o tamanho do mesmo.

Além de sofrerem com a "maldição da alta dimensionalidade", os sistemas CBIR também sofrem com o problema do "gap semântico", que é causado pela inconsistência que existe entre as características de baixo-nível automaticamente extraídas das imagens e sua interpretação humana de alto-nível. Assim, é importante encontrar as características que melhor representam o conteúdo semântico da imagem.

Este capítulo discute como aplicar técnicas de mineração de regras de associação estatísticas para melhorar a recuperação de imagens por conteúdo. Aqui, é apresentado o algoritmo StARMiner (***S**tatistical **A**ssociation **R**ule **M**iner*), desenvolvido para atuar na tarefa de redução de dimensionalidade dos vetores de características, lidando com a maldição da alta dimensionalidade que degrada a recuperação de imagens por conteúdo. O algoritmo usa medidas estatísticas para descrever o comportamento das características considerando as categorias (classes)

das imagens e para encontrar regras de associação representativas. As regras encontradas pelo algoritmo StARMiner, também são aplicadas para ponderar características dos vetores e para melhorar a precisão das consultas. A técnica de mineração aqui apresentada também visa reduzir o "gap semantico", pois ela encontra associações entre as características de baixo nível e características de alto-nível das imagens.

5.2 Descrição do Algoritmo StARMiner

A mineração de regras de associação é uma tarefa destinada a encontrar relacionamentos entre dados em uma base de dados. Devido a sua simplicidade, facilidade de entendimento e por refletir a maneira como o ser humano adquire novos conhecimento através de associações, a mineração de regras de associação tem sido uma das tarefas mais estudas e pesquisadas da literatura, tendo uma vasta aplicabilidade.

As características extraídas das imagens são valores numéricos contínuos. Vetores de características descrevem as imagens quantitativamente. Assim, uma técnica apropriada para encontrar regras de associação envolvendo imagens deve considerar dados quantitativos. Infelizmente, a definição de regras de associação categóricas (ver capítulo 4) não é diretamente mapeada para o caso de atributos quantitativos.

Embora várias técnicas de discretização tenham sido propostas e discutidas no capítulo 4 para diminuir os problemas referentes a discretização dos dados, a primeira abordagem do trabalho desenvolvido nesta tese foi trabalhar com regras de associação empregando diretamente os dados contínuos, sem discretizá-los.

Nesta seção é apresentado o algoritmo StARMiner (*Statistical Association Rule Miner*), um novo algoritmo para mineração de regras de associação estatísticas. O objetivo do algoritmo StARMiner é encontrar regras de associação que selecione o conjunto mínimo de características que preserve a habilidade de diferenciar imagens de acordo com suas categorias.

Seja x_j uma categoria (classe) e f_i uma característica (atributo) de uma imagem. As regras retornadas pelo algoritmo StARMiner têm o formato geral:

$x_j \rightarrow f_i$,

onde o antecessor da regra indica um subconjunto de imagens que pertencem a categoria x_j e o sucessor da regra f_i é uma característica que tem um comportamento diferente em imagens da categoria x_j em relação às demais imagens da base de dados. O algoritmo StARMiner somente retorna regras que satisfazem as duas condições a seguir:

Condição 5.1 *A característica f_i deve ter um comportamento em imagens da categoria x_j diferente do seu comportamento em imagens das demais categorias da base.*

Condição 5.2 *A característica f_i deve apresentar um comportamento uniforme nas imagens da categoria x_j.*

As condições 5.1 e 5.2 são implementadas no algoritmo StARMiner através da incorporação de restrições de interesse para a mineração das regras. Essas restrições de interesse são descritas a seguir. Seja T uma base de imagens médicas, x_j uma categoria de uma imagem, $T_{x_j} \in T$ o subconjunto de imagens da categoria x_j, f_i a i-ésima característica do vetor de características F, e f_{i_k} o valor da característica f_i na imagem k.

Sejam $\mu_{f_i}(V)$ e $\sigma_{f_i}(V)$, respectivamente, a média e o desvio padrão dos valores da característica f_i no subconjunto de imagens V. O algoritmo usa três limiares definidos pelo usuário:

- $\Delta\mu_{min}$ - a mínima diferença das médias da característica f_i entre as imagens da categoria x_j e as demais imagens da base;
- σ_{max} - o máximo desvio padrão permitido da característica f_i em imagens da categoria x_j;
- γ_{min} - a mínima confiança para rejeitar a hipótese $H0$, de que são iguais estatisticamente às médias dos valores de f_i nos conjuntos T_{x_j} (imagens da categoria x_j) e $T - T_{x_j}$ (imagens das demais categorias).

O StARMiner minera regras da forma $x_j \rightarrow f_i$, se as condições fornecidas nas equações 5.3, 5.4 and 5.5 forem satisfeitas.

$$\mu_{f_i}(V) = \frac{\sum_{k \in V}(f_{i_k})}{|V|} \tag{5.1}$$

$$\sigma_{f_i}(V) = \sqrt{(\frac{\sum_{k \in V}(f_{i_k} - \mu_{f_i}(V))^2}{|V|})} \tag{5.2}$$

$$\mu_{f_i}(T_{x_j}) - \mu_{f_i}(T - T_{x_j}) \geq \Delta\mu_{min} \tag{5.3}$$

$$\sigma_{f_i}(T_{x_j}) \leq \sigma_{max} \tag{5.4}$$

$$\text{Rejeição da Hipótese } H_0 : \mu_{f_i}(T_{x_j}) = \mu_{f_i}(T - T_{x_j}) \tag{5.5}$$

Na equação 5.5, a hipótese H_0 deve ser rejeitada com a confiança igual ou maior do que γ_{min}, em favor da hipótese de que as médias $\mu_{f_i}(T_{x_j})$ e $\mu_{f_i}(T - T_{x_j})$ são estatisticamente diferentes. Para rejeitar H_0 com confiança γ_{min}, o valor de Z, calculado usando a equação 5.6, deve estar na região de rejeição ilustrada na figura 5.1. Os valores críticos de Z, que são Z_1 e Z_2, dependem do valor de γ_{min} como é apresentado na tabela 5.1:

$$Z = \frac{\mu_{f_i}(T_{x_j}) - \mu_{f_i}(T - T_{x_j})}{\frac{\sigma_{f_i}(T_{x_j})}{\sqrt{(|T_x|)}}} \tag{5.6}$$

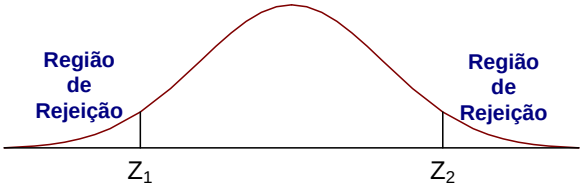


Figura 5.1: Ilustração das regiões de rejeição do teste de hipóteses.

Tabela 5.1: Exemplos de valores críticos de Z.

γ_{min}	**0.9**	**0.95**	**0.99**
Z_1	-1.64	-1.96	-2.58
Z_2	1.64	1.96	2.58

Uma regra $x_j \rightarrow f_i$ retornada pelo algoritmo, relaciona uma característica f_i com uma categoria x_j, onde os valores de f_i tem um comportamento estatisticamente diferente em imagens da categoria x_j. Esta propriedade indica que f_i é uma característica importante para distinguir imagens da categoria x_j das demais imagens.

O algoritmo StARMiner também fornece informação sobre o comportamento das características presentes em suas regras. Uma regra minerada pelo algoritmo StARMiner, em sua forma completa é:

$$x_j \rightarrow f_i,\ \mu_{f_i}(T_{x_j}),\ \mu_{f_i}(T - T_{x_j}),\ \sigma_{f_i}(T_{x_j}),\ \sigma_{f_i}(T - T_{x_j})$$

onde, $\mu_{f_i}(T_{x_j})$ e $\sigma_{f_i}(T_{x_j})$ são, respectivamente, a média e o desvio padrão da característica f_i nas imagens da categoria x_j; $\mu_{f_i}(T - T_{x_j})$ e $\sigma_{f_i}(T - T_{x_j})$ são, respectivamente, a média e o desvio padrão dos valores de f_i em imagens que não são da categoria x_j.

O StARMiner é apresentado no algoritmo 2. O algoritmo StARMiner realiza duas varreduras na base de dados. Na primeira varredura, o valor de média de cada característica é calculado (linhas 1 a 6). Na segunda varredura (linhas 7 a 16), o valor do desvio padrão e o valor de Z para cada característica é calculado. As restrições de interesse são processadas nas linhas 11 e 12. Uma regra é retornada somente se ele satisfizer os limiares $\Delta\mu_{min}$, σ_{max} e γ_{min} fornecidos como parâmetros de entrada do algoritmo. A complexidade do algoritmo StARMiner é $\Theta(ckN)$, onde N é o número de instâncias da base, k é o número de características, e c é o número de categorias.

Algoritmo 2: Algoritmo StARMiner

Dados: Base de dados T de tuplas de imagens estruturadas como $\{x_j, f_1, f_2,...,f_n\}$ onde x_j representa a categoria da imagem e f_i uma característica da imagem; limiares $\Delta\mu_{min}$, σ_{max} e γ_{min}

Resultado: As regras mineradas

```
1  Percorra a base de dados T ;
2  para cada característica f_i faça
3      para cada categoria x_j faça
4          calcule μ_{f_i}(T_{x_j}) e μ_{f_i}(T − T_{x_j}) ;
5      fim
6  fim
7  Percorra a base de dados T;
8  para cada característica f_i faça
9      para cada categoria x_j faça
10         calcule σ_{f_i}(T_{x_j}) e σ_{f_i}(T − T_{x_j}) ;
11         calcule o valor Z_{i_j} ;
12         se (μ_{f_i}(T_{x_j}) − μ_{f_i}(T − T_{x_j})) ≥ Δμ_min e σ_{f_i}(T_{x_j}) ≤ σ_max e (Z_{i_j} < Z_1 ou Z_{i_j} > Z_2)
           então
13             escreva x_j → f_i, μ_{f_i}(T_{x_j}), μ_{f_i}(T − T_{x_j}), σ_{f_i}(T_{x_j}), σ_{f_i}(T − T_{x_j}) ;
14         fim
15     fim
16 fim
```

5.2.1 Seleção de Características usando o algoritmo StARMiner

Uma importante questão é como utilizar o conhecimento obtido através das regras de associação estatísticas para a seleção de características. O algoritmo StARMiner encontra regras que revelam quais as características têm o maior poder de diferenciação das categorias das imagens. Isso porque o algoritmo minera regras que envolvem características que apresentam um comportamento uniforme e particular em imagens de uma determinada categoria. Isso é importante pois, as características (ou atributos) que apresentam um comportamento uniforme para todas as imagens na base de dados, independente da categoria da imagem, não contribuem para categorizá-las e devem ser eliminadas do vetor de características. A condição 5.3, apresentada a seguir, define como é feito o processo de seleção de características binária a partir das regras de associação mineradas pelo algoritmo StARMiner.

Condição 5.3 *As características presentes no conjunto de regras retornadas pelo algoritmo StARMiner são selecionadas como as mais relevantes.*

Conforme ilustraremos na seção de experimentos, a seleção binária de características feita utilizando as regras de associação estatísticas é bastante adequada para reduzir a dimensionalidade de bases de imagens médicas.

Além da seleção binária de características, as regras de associação mineradas pelo algoritmo StARMiner também servem para a seleção de características contínua, que corresponde à ponderação do vetor de característica dando um maior peso para as características que mais discriminam as imagens em categorias. Supondo que as imagens são classificadas em m categorias $X = x_1, x_2,..., x_m$, para cada característica f_i, o algoritmo StARMiner busca por regras da forma $x_j \rightarrow f_i$. Isto é, o algoritmo StARMiner busca relacionar cada característica f_i com cada categoria x_j. Se uma regra $x_j \rightarrow f_i$ é encontrada, significa que a característica f_i discrimina bem as imagens da classe x_j. Assim, as características f_i que melhor discriminam as imagens são aquelas que geram regras $x_j \rightarrow f_i$, para todo $x_j \in X$, ou seja, elas discriminam bem as imagens de todas as categorias. Do mesmo modo, as características menos discriminantes são aquelas que não geram nenhuma regra, significando que elas têm um comportamento uniforme em todas as categorias da base, não sendo úteis para distinguir imagens de diferentes categorias. Assim, para ponderar uma característica f_i, é utilizado o número de regras em que f_i ocorre. As equações 5.7 e 5.8 são utilizadas para promover a ponderação dos vetores de características.

$$w_i = 10r_i + 1 \tag{5.7}$$

$$w_i = 10r_i \tag{5.8}$$

Nas equações 5.7 e 5.8, r_i é o número de regras mineradas em que f_i aparece. Essas equações foram obtidas empiricamente. A diferença entre as equações 5.7 e 5.8 é que na primeira não ocorre a redução de dimensionalidade do vetor de característica, enquanto na segunda ocorre a redução de dimensionalidade (as características f_i que não ocorrem em nenhuma regra tem peso $w_i = 0$ e são eliminadas).

5.3 Experimentos

O algoritmo StARMiner tem sido utilizado em uma grande gama de trabalhos com diferentes aplicações:

- (a) seleção binária de características [Ribeiro et al., 2005a] [Ribeiro et al., 2006d] [Silva et al., 2008];

- (b) seleção binária de características em conjunto com o algoritmo FD-ASE [Sousa et al., 2002], que utiliza conceitos de dimensão fractal para correlacionar atributos (características), permitindo uma maior redução de dimensionalidade dos conjuntos de dados [Felipe et al., 2006];

- (c) em procedimentos de realimentação de relevância para maximizar a precisão das buscas por conteúdo [Ribeiro et al., 2006c];

- (d) ponderação de vetores de características para maximizar a precisão das buscas por conteúdo [Bugatti et al., 2008].

Os principais estudos de caso relativos à aplicação do algoritmo StARMiner são detalhados nesta seção. Em geral, os estudos de caso validam o algoritmo StARMiner através de um procedimento composto de 3 passos ilustrados na Figura 5.2.

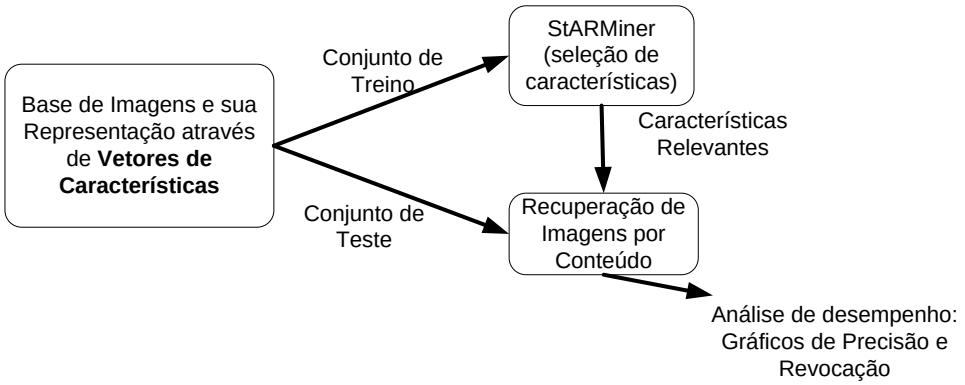


Figura 5.2: Passos do procedimento utilizado para validar o algoritmo StARMiner.

Observe a figura 5.2. Como o StARMiner é um algoritmo supervisionado de seleção de características, o conjunto de imagens é dividido em conjunto de treino e de teste. O conjunto de treino é submetido ao algoritmo de seleção de características, o conjunto de teste é utilizado para avaliar o desempenho do método de redução de dimensionalidade nas consultas por conteúdo. Os vetores de características das imagens de treino são submetidos ao algoritmo de seleção de características. Depois, os vetores reduzidos são utilizados para indexar as imagens do conjunto de teste para a realização das consultas por conteúdo, reduzindo o custo computacional da execução das consultas. Para cada imagem do conjunto de teste, várias consultas aos k-vizinhos mais próximos são executadas (variando k de 1 até o tamanho da base de teste). As medidas de precisão e revocação (P&R) são computadas para cada cada resultado de consulta e a curva média de P&R é computada para ambos os vetores de características: o original e o reduzido. Uma regra geral para a análise dos gráficos de precisão e revocação P&R é: quanto mais próxima estiver a curva do topo do gráfico, melhor é a técnica.

5.3.1 Bases de Imagens

Uma série de bases de imagens médicas foi utilizada para realizar os experimentos. Para cada tipo de imagem médica e para cada tipo de análise, um determinado vetor de características foi

utilizado para a sua representação. Aqui são descritas as bases de imagens médicas utilizadas para validar os métodos propostos. Para cada base, são descritos os tipos de imagens da base, o processamento aplicado sobre essas imagens e os vetores de características utilizados para representá-las. Alguns vetores de características utilizados são apontados na literatura como os mais adequados para representar uma determinada base. Outros são escolhidos em decorrência de resultados de experimentos realizados durante o trabalho desta tese.

Para facilitar o entendimento foi adotada uma formalização para a definição dos nomes das bases que é composto de três partes: (a) parte do nome do extrator utilizado para a representação da base; (b) parte de um nome que descreve o tipo das imagens da base; e (c) número de imagens da base.

Base BalanRMI704

A base BalanRMI704 é formada por 704 imagens de Ressonância Magnética (RM) obtidas do Hospital das Clínicas da USP em Ribeirão Preto. As imagens dessa base são classificadas nas 8 categorias apresentadas na tabela 5.2.

Tabela 5.2: Configuração da base de imagens BalanRMI704.

Categoria	**# de imagens**
Angiograma	36
Pélvis Axial	86
Cabeça Axial	155
Cabeça Sagital	258
Abdômen Coronal	23
Espinha Sagital	59
Abdômen Axial	51
Cabeça Coronal	36
Total	**704**

A base de dados foi previamente segmentada. O método proposto por Balan [Balan, 2007], apresentado no capítulo 2 (seção 2.2.4), foi utilizado para segmentar as imagens da base BalanRMI704. A segmentação da base foi realizada considerando 5 regiões de segmentação, que é a configuração que produz os maiores valores de precisão para esta base. Para cada região segmentada, seis características foram extraídas: a massa m ou tamanho; o centro de massa ou centróide (xc, yc); o nível de cinza médio a; a dimensão fractal (dimensão de correlação) D_2 e o coeficiente linear b usado para estimar D_2. Essas características são apontadas em [Balan, 2007] como as mais representativas para essa base. Assim, quando a imagem é segmentada em 5 classes, o vetor de características tem $5 * 6 = 30$ elementos. A figura 5.3 ilustra o vetor de característica empregado para representar a base BalanRMI704.

É importante enfatizar que considerando apenas 5 regiões da imagem (como o ilustrado na figura 5.3), o vetor de características gerado já é um vetor compacto (30 elementos). Esse vetor

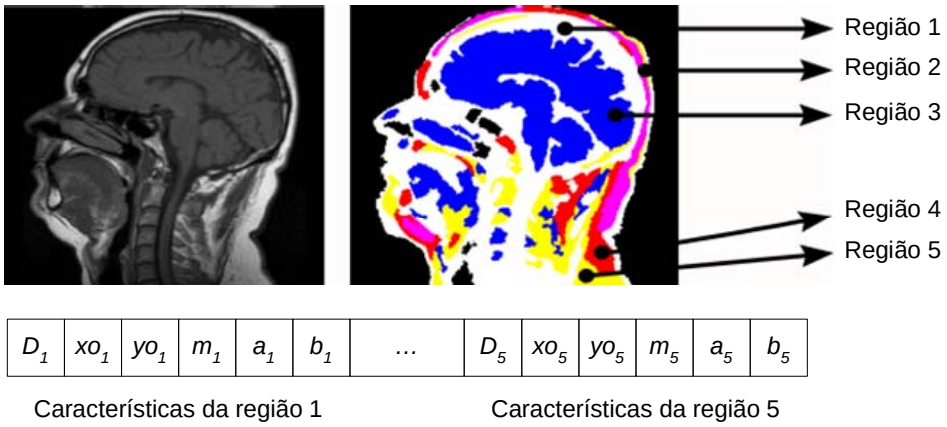
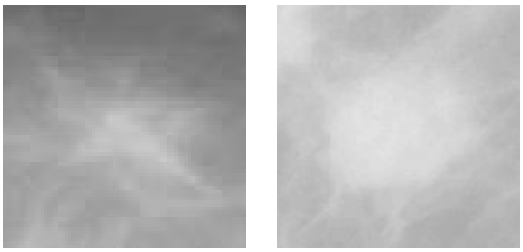


Figura 5.3: Regiões de segmentação e o vetor de características utilizado para representar a base BalanRMI704.

compacto discrimina bem as imagens, no entanto, a aplicação do algoritmo StARMiner mostra que ele ainda contém informação supérflua que não necessita ser armazenada.

Base ZernikeMamo250

A base ZernikeMamo250 consiste de 250 imagens obtidas da biblioteca *Digital Database for Screening Mammography* - DDSM [Heath et al., 2001]. As imagens são ROIs (Região de Interesse) compreendendo massas tumorais, tiradas de mamografias. As imagens são classificadas como benignas ou malignas, de acordo com uma análise feita anteriormente por radiologistas e eventualmente confirmada por exames complementares. Exemplos de imagens da base são exibidos na figura 5.4.



(a) (b)

Figura 5.4: Atividade do Exemplos de imagens da base ZernikeMamo250: massa maligna (a) e massa benigna (b).

O vetor de características utilizado para representar essa base compreende os polinômios de Zernike de ordem 30, compondo um vetor de características de 256 elementos para representar cada imagem. Os polinômios de Zernike foram discutidos no capítulo 2 (seção 2.3.4).

Base TexturaRMI943

A base de dados TexturaRMI943 consiste em 943 imagens médicas obtidas do Hospital

das Clínicas da USP em Ribeirão Preto. A base de dados TexturaRMI943 é classificada em 6 categorias: abdômen coronal, angiograma, cabeça axial, cabeça coronal, cabeça sagital e espinha sagital. Um exemplo de cada tipo de imagem da base TexturaRMI943 é apresentada na figura 5.5.

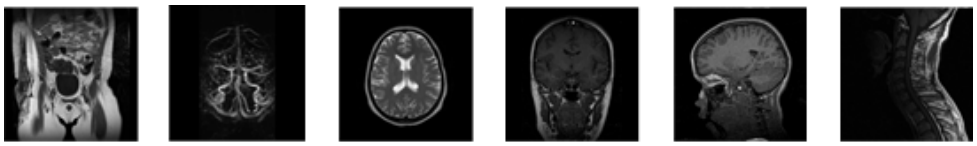


Figura 5.5: Um exemplo de cada tipo de imagem da base TexturaRMI943.

As características de textura propostas em [Haralick et al., 1973] foram extraídas e organizadas em vetores de características. Para realizar a extração, primeiramente os níveis de cinza das imagens foram reduzidos para 16. Uma matriz de co-ocorrência (discutida na seção 2.3.5) foi gerada para cada imagem, para as direções de 0°, 45°, 90° e 135°, e para as distâncias 1, 2, 3, 4 e 5. Assim, vinte matrizes de 16×16 elementos inteiros por imagem foram produzidas. Para cada matriz, os sete descritores descritos na tabela 5.3 foram calculados, produzindo um vetor de características de 140 elementos para representar cada imagem. Esse procedimento para a extração de características é detalhado em [Felipe et al., 2003].

Tabela 5.3: Características de textura e suas posições no vetor de característica.

nome	**equação**	**significado**	**posição**
Step	$\sum_i \sum_j P(i,j)$	distribuição	1-20
Variância	$\sum_i \sum_j (i-j)^2 P(i,j)$	contraste	21-40
Entropia	$\sum_i \sum_j P(i,j) log(P(i,j))$	suavidade	41-60
Energia	$\sum_i \sum_j P(i-j)^2$	uniformidade	61-80
Homogeneidade	$\sum_i \sum_j \frac{P(i-j)}{(1+\|i-j\|)}$	homogeneidade	81-100
3° Momento	$\sum_i \sum_j (i-j)^3 P(i,j)$	distorção	101-120
Inv. Variância	$\sum_i \sum_j \frac{P(i,j)}{(i-j)^2}$	contraste inverso	121-140

Base TexturaRMI704

A base TexturaRMI704 é formada pelas mesmas 704 imagens médicas da base Balan704RMI, porém descritas pelo vetor de características composto das 140 características de textura conforme realizado para a base TexturaRMI943.

A implementação utilizada dos algoritmos Relief-F e DTM foi obtida na ferramenta Weka (www.cs.waikato.ac.nz/ml/weka). Os parâmetros do Relief-F são $knn = 10$ (10 vizinhos mais próximos são avaliados para cada instância), $sigma = 2$ (parâmetro que indica o quão rápido o peso dos atributos diminui com a distância. A configuração do algoritmo DTM foi $nmif = 2$ (número mínimo de instância por folha) e $conf = 2$ (fator de confiança para realizar podas na

árvore de decisão). A implementação do algoritmo StARMiner foi feita utilizando C++ padrão. Os parâmetros utilizados como entrada do algoritmo StARMiner estão descritos nos estudos de caso.

5.3.2 Estudo de Caso 1

Neste estudo de caso o algoritmo StARMiner é utilizado para determinar um conjunto mínimo e representativo de características da base de dados de acordo com a condição 5.3, onde as características presentes nas regras retornadas pelo algoritmo StARMiner são selecionadas como as mais relevantes e as que não estão presentes nas regras são removidas.

A base de dados BalanRMI704 foi dividida em: conjunto de treino, composto de 176 imagens; e conjunto de teste, composto de 528 imagens. O algoritmo StARMiner foi aplicado sobre os vetores de características (compostos de 30 características) das imagens de treino, gerando 21 regras. Foram avaliados vários valores de limiares e os melhores resultados foram conseguidos usando $\Delta\mu_{min} = 0.2$, $\sigma_{max} = 0.13$ e $\gamma_{min} = 0.98$. Um exemplo de regra obtida é:

$$\textit{angiograma} \rightarrow \textit{ média de nível de cinza na região 2 } (a_2)$$
$$\mu(a_2)\text{(imagens de angiograma) = 0.1}$$
$$\mu(a_2)\text{(imagens que não são angiogramas) = 0.43}$$
$$\sigma(a_2)\text{(imagens de angiograma) = 0.07}$$
$$\sigma(a_2)\text{(imagens que não são angiogramas) = 0.18}$$

Essa regra indica a importância da média do nível de cinza da região 2 (a_2) para identificar imagens de angiograma. Essa regra mostra que, a média de a_2 em imagens de angiograma (0.1) é diferente da média dessa mesma característica para as imagens restantes (0.43). O desvio padrão de a_2 em imagens de angiograma é pequeno (0.07) quando comparando com o desvio padrão de valores de a_2 para as imagens restantes (0.18). Esses valores indicam que a característica a_2 (média de nível de cinza da região 2) tem um comportamento particular em imagens de angiograma, contrastando com seu comportamento em imagens de outras categorias, evidenciando que é uma característica relevante para distinguir as imagens de angiograma.

A tabela 5.4 mostra as características selecionadas pelo StARMiner. As características selecionadas estão *sublinhadas*. As características não sublinhadas devem ser eliminadas do vetor de características.

Os resultados indicam que a dimensão fractal D tem uma pequena contribuição para distinguir imagens desta base (veja a tabela 5.4, linha 1). Os resultados também indicam que a média de nível de cinza a é uma das características mais importantes para representar as imagens, pois nenhuma característica deste tipo foi removida do vetor de características.

O algoritmo StARMiner selecionou 21 características neste estudo de caso e a habilidade do mesmo na tarefa de seleção de características foi avaliada comparativamente. Para isso, o algoritmo Relief-F também foi aplicado aos dados de treino. As 21 características mais relevantes

Tabela 5.4: Características selecionadas pelo algoritmo StARMiner. As características selecionadas estão *sublinhadas*. As características não sublinhadas devem ser eliminadas do vetor de características.

Região 1	*Região 2*	*Região3*	*Região 4*	*Região 5*
D	*D*	*D*	*D*	*D*
xo	*xo*	***xo***	***x0***	***xo***
yo	***yo***	***yo***	***yo***	***yo***
m	***m***	***m***	*m*	*m*
a	***a***	***a***	***a***	***a***
b	***b***	***b***	***b***	***b***

retornadas pelo algoritmo Relief-F foram utilizadas para compor um vetor de características. O algoritmo DTM também foi aplicado nas imagens de treino e as 21 características mais relevantes retornadas foram selecionadas para compor um vetor de características.

Para construir os gráficos de Precisão e Revocação (*Precision vs. Recall* - P&R), foram considerados quatro casos de vetores de características para representar as imagens:

- (a) usando o vetor original com 30 características;
- (b) usando as 21 características selecionadas pelo algoritmo StARMiner;
- (c) usando as 21 características selecionadas pelo algoritmo Relief-F;
- (d) usando as 21 características selecionadas pelo algoritmo DTM.

Consultas por similaridade foram executadas sobre o conjunto de teste e o gráfico de P&R foi construído. A Figura 5.6 mostra o gráfico de P&R obtido.

O gráfico da Figura 5.6 mostra que os resultados obtidos com 21 características são melhores do que os resultados obtidos usando todas as 30 características originais. Assim, embora usando aproximadamente 70% do esforço original de processamento requerido e demandando menos memória, a precisão das buscas por conteúdo foi melhorada. O esforço computacional de uma consulta por similaridade é proporcional ao tamanho do vetor de características empregado para representar as imagens.

Para garantir que foi selecionado o conjunto mínimo de características relevantes que mantém a precisão dos resultados, também foram executadas as mesmas consultas por similaridade usando 20 características, removendo randomicamente uma das 21 selecionadas pelo algoritmo StARMiner, e , como resultado disso, a precisão sempre diminui. A figura 5.7 mostra um gráfico de P&R comparando a precisão das consultas usando todas as 30 características, as 21 selecionadas pelo StARMiner e as 20 obtidas removendo randomicamente uma das selecionadas pelo StARMiner.

As características selecionadas mostradas na tabela 5.4 indicam que a característica *D* tem uma pequena contribuição para representar as imagens. As consultas por similaridade foram

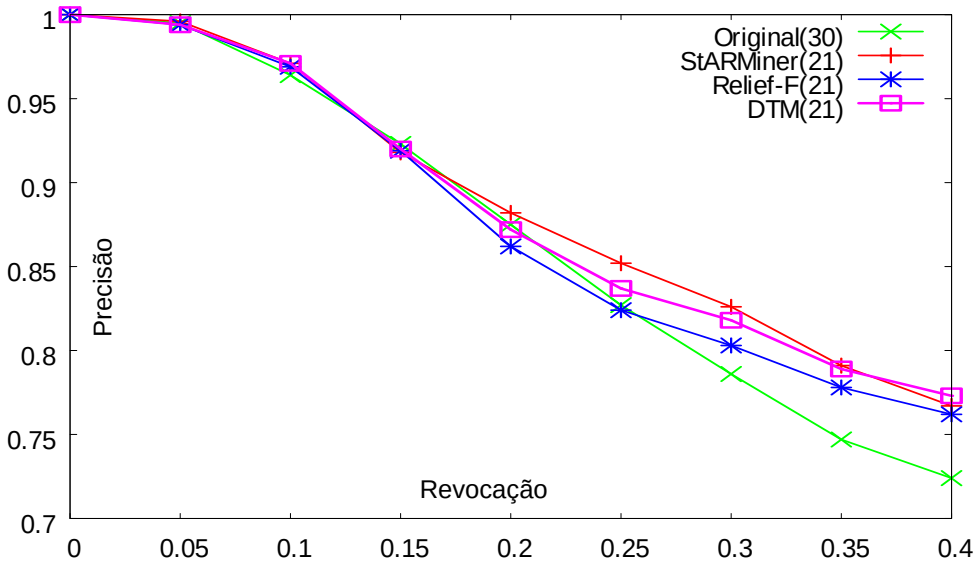


Figura 5.6: Gráfico de P&R construído usando o conjunto de teste da base BalanRMI704 representado por: as 30 características originais, as 21 selecionadas pelo algoritmo StARMiner, as 21 selecionadas pelo algoritmo Relief-F e as 21 selecionadas pelo algoritmo DTM.

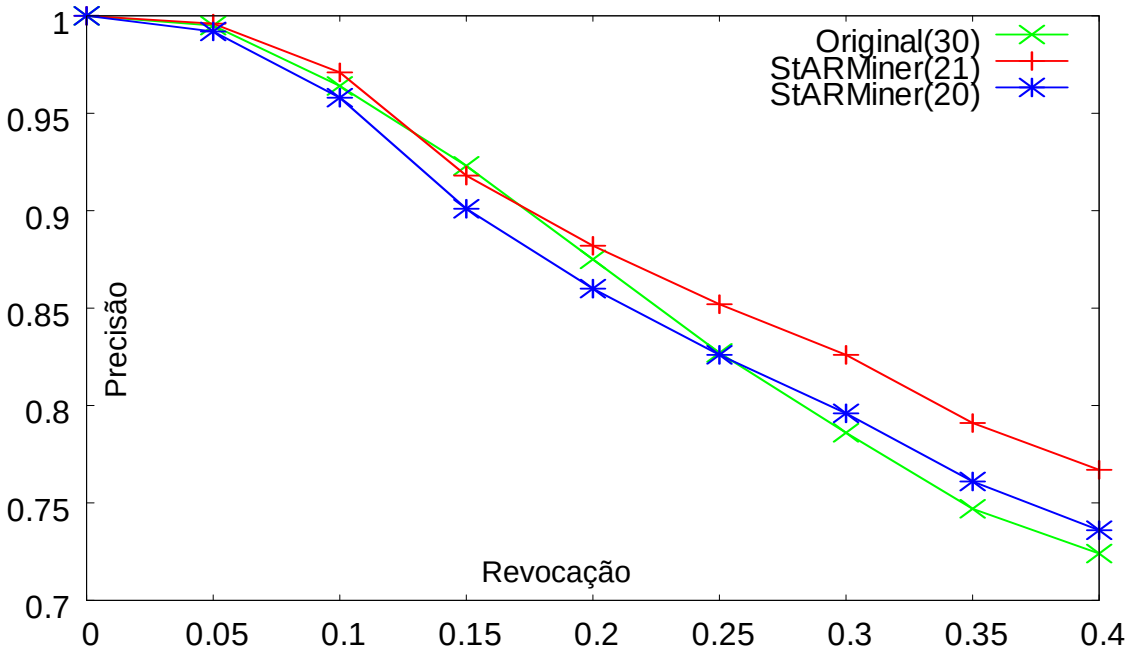


Figura 5.7: Gráfico de P&R obtido usando as 30 características originais, as 21 características selecionadas pelo StARMiner e as 20 características obtidas removendo randomicamente uma característica do conjunto das selecionadas pelo StARMiner.

refeitas removendo a característica *D* da região 1 do conjunto de 21 características selecionadas pelo StARMiner para checar se os valores de precisão diminuem. Os resultados deste teste são apresentados na figura 5.8.

Comparando as curvas apresentadas na figura 5.8, é possível notar que removendo a ca-

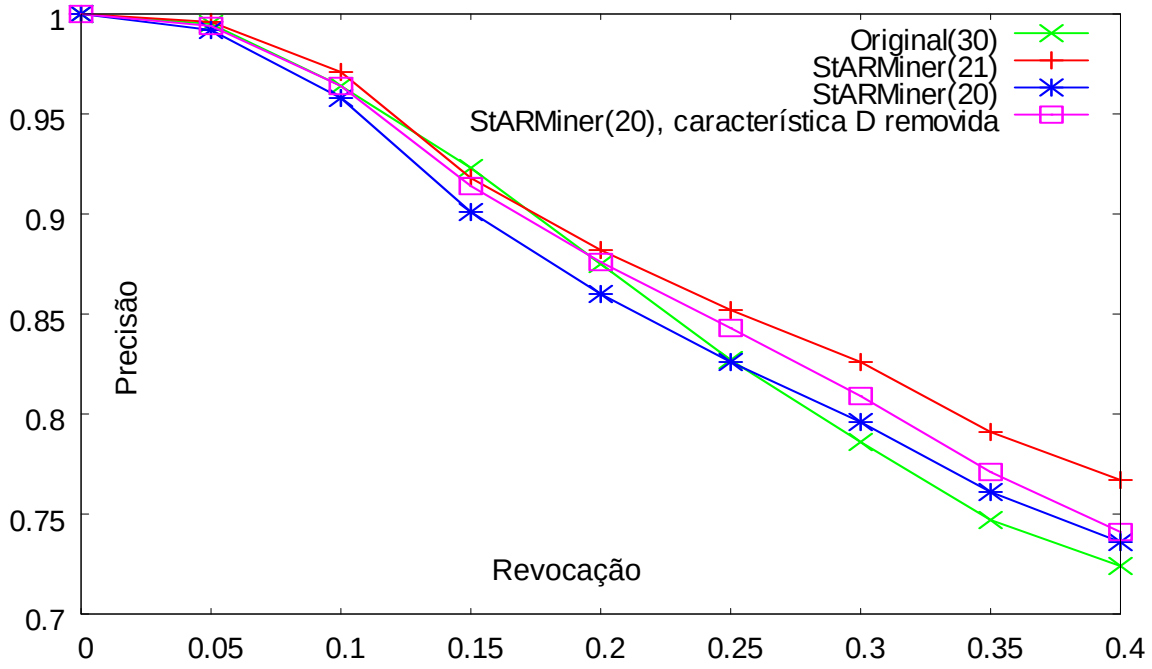


Figura 5.8: Gráfico de P&R obtido usando as 30 características originais, as 21 características selecionadas pelo StARMiner, as 20 características obtidas removendo randomicamente uma característica do conjunto selecionado pelo StARMiner e as 20 características obtidas removendo a característica *D* do conjunto das selecionadas pelo StARMiner.

racterística *D* do conjunto das 21 selecionadas pelo algoritmo StARMiner, os valores de precisão reduzem menos do que removendo outra característica, pois a característica *D* tem uma contribuição pequena para a discriminação das imagens. Para a região 1, a característica *D* é significativa, entretanto, comparando com as demais característica selecionadas, ela é a que traz a menor contribuição para diferenciar os diferentes tipos de imagens.

É interessante visualizar as imagens retornadas em uma consulta por similaridade usando o conjunto de todas as características e o conjunto de características reduzido para representar as imagens. Foram executadas duas consultas aos 15-vizinhos mais próximos sobre a mesma imagem de consulta apresentada no canto superior esquerdo da tela das figuras 5.9 e 5.10

A figura 5.9 mostra os resultados usando o vetor original de 30 características, enquanto a figura 5.10 mostra os resultados usando apenas as 21 selecionadas pelo StARMiner. As características selecionadas pelo StARMiner produzem um melhor resultado do que as 30 originais para a consulta executada. A consulta executada usando as 30 características (figura 5.9) alcançou uma precisão de 80%, enquanto a consulta executada usando as 21 características selecionadas atingiu uma precisão de 100% (figura 5.10). Isso aconteceu em outras situações, comprovando a afirmação anterior de que a presença de características correlacionadas no vetor de características, ao invés de melhorar, deteriora o processo de recuperação de imagens por conteúdo.

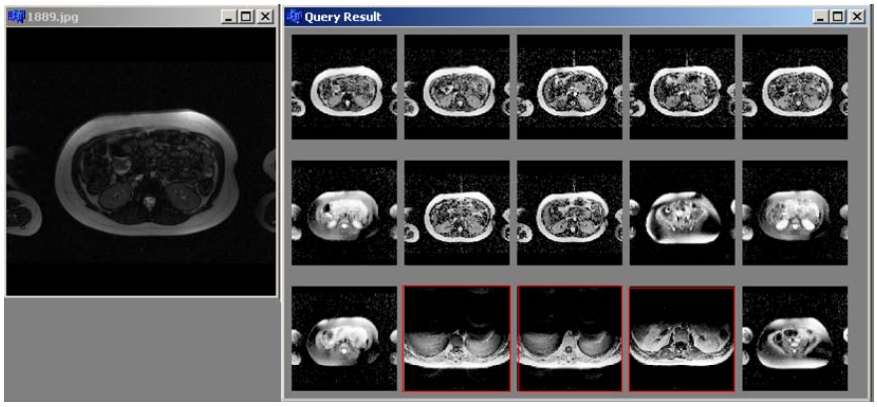


Figura 5.9: Exemplo de consulta usando as 30 características originais.

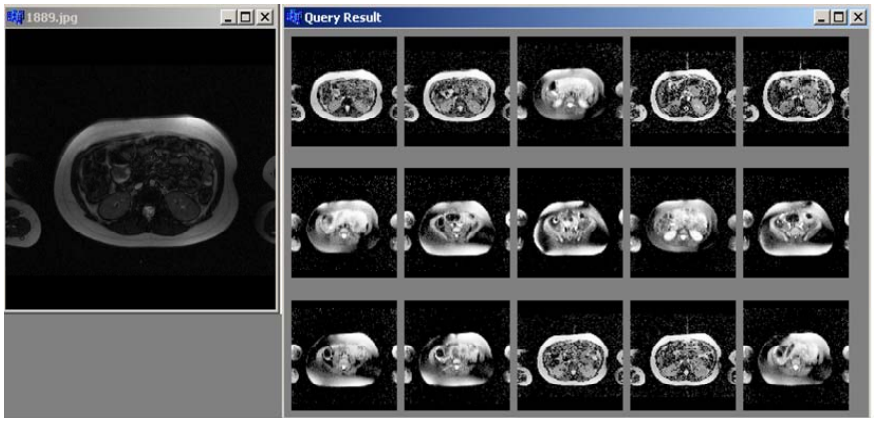


Figura 5.10: Exemplo de consulta usando as 21 características selecionadas pelo StARMiner.

5.3.3 Estudo de Caso 2

Neste estudo de caso o algoritmo StARMiner é utilizado para determinar um conjunto mínimo e representativo de características da base de dados. A base de dados ZernikeMamo250, composta de 250 imagens de ROIs, descrita no começo desta seção, foi utilizada. Da base ZernikeMamo250, 89 imagens foram utilizadas para treino.

O algoritmo DTM (*Decision Tree Method*) foi aplicado sobre os vetores de características (compostos de 256 características) das imagens de treino, gerando uma árvore de decisão com 38 características. O conjunto de imagens de treino também foi submetido ao algoritmo StARMiner, que foi calibrado para selecionar um conjunto de 38 características (o mesmo número de características selecionadas pelo DTM), no entanto, selecionando um conjunto de características diferente do conjunto selecionado pelo DTM. Um exemplo de regra obtida pelo STARMiner é:

$$\textit{tumor maligno} \rightarrow \ 9^{\circ}\ \textit{Momento de Zernike}\ (M_9)$$
$$\mu(M_9)\text{(imagens de tumor maligno)} = 0.3$$
$$\mu(M_9)\text{(imagens que não são de tumor maligno)} = 0.1$$
$$\sigma(M_9)\text{(imagens de tumor maligno)} = 0.1$$

$\sigma(M_9)$(imagens que não são de tumor maligno) = 0.28

A regra acima indica a relevância do 9° Momento de Zernike em diferenciar tumores malignos e benignos. A figura 5.11 mostra o gráfico de P&R construído usando o conjunto Zernike-Mamo250 representado pelas 256 características originais, pelas 38 características selecionadas pelo StARMiner, e pelas 38 selecionadas pelo algoritmo DTM.

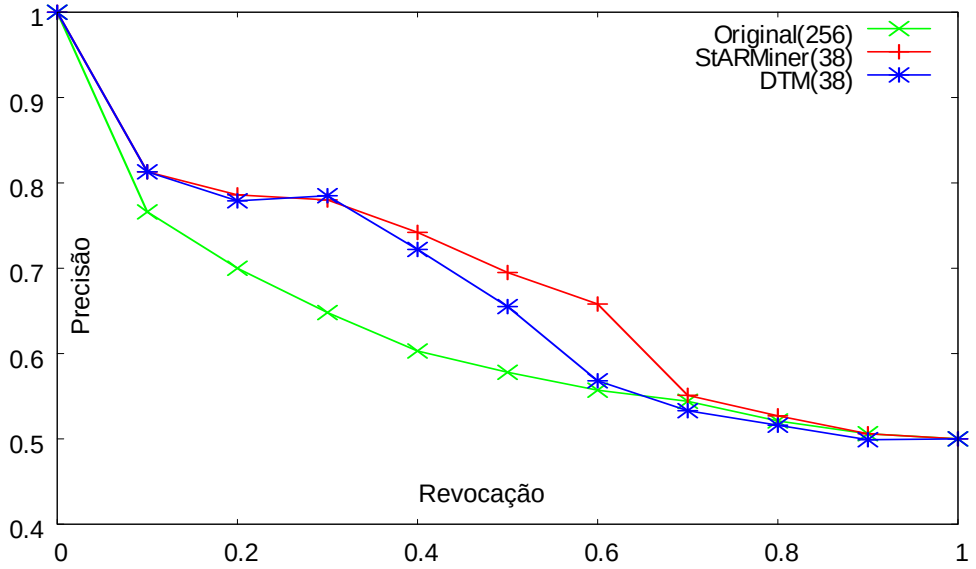


Figura 5.11: Curvas de P&R obtidas utilizando o conjunto ZernikeMamo250 representado por: (a) as 256 características originais; (b) as 38 características selecionadas pelo StARMiner; e (c) as 38 selecionadas pelo algoritmo DTM.

Analisando o gráfico da figura 5.11, é possível concluir que os resultados obtidos usando 38 características são melhores do que utilizando 256 características para ambos os algoritmos, StARMiner e DTM. Assim, a redução de dimensionalidade providencia um importante ganho de precisão dentro da região de revocação compreendendo os valores menores que 70%. As regiões com os valores baixos de revocação são as mais importantes nos ambientes de CBIR, pois na execução de uma consulta aos k-vizinhos mais próximos usualmente não é utilizado grandes valores de k. Esses resultados ilustram que a "maldição da alta-dimensionalidade" realmente deturpa os resultados das consultas. O gráfico da figura 5.11 mostra que o StARMiner alcança maiores valores de precisão do que o DTM para as regiões de revocação entre 30% e 70%.

5.3.4 Estudo de Caso 3

Neste estudo de caso o algoritmo StARMiner é utilizado para determinar um conjunto mínimo e representativo de características da base TexturaRMI943. A base TexturaRMI943 foi divida em conjunto de treino (composto de 376 imagens) e conjunto de teste (composto de 567 imagens).

O algoritmo StARMiner foi executado sobre as imagens de treino, gerando 100 regras. Os parâmetros utilizados pelo algoritmo foram: $\Delta\mu_{min} = 0.1$, $\sigma_{max} = 0.2$, $\gamma_{min} = 0.9$. O algoritmo StARMiner selecionou 100 características e essas 100 características foram utilizadas para compor os vetores de características para representar as imagens. O algoritmo DTM foi executado, onde as 100 características mais relevantes foram utilizadas para compor um outro conjunto de vetores de características para representar as imagens. A figura 5.12 foi obtida usando 3 tipos de vetores de características:

- (a) as 140 características originais;
- (b) as 100 características selecionadas pelo StARMiner;
- (c) as 100 características selecionadas pelo DTM.

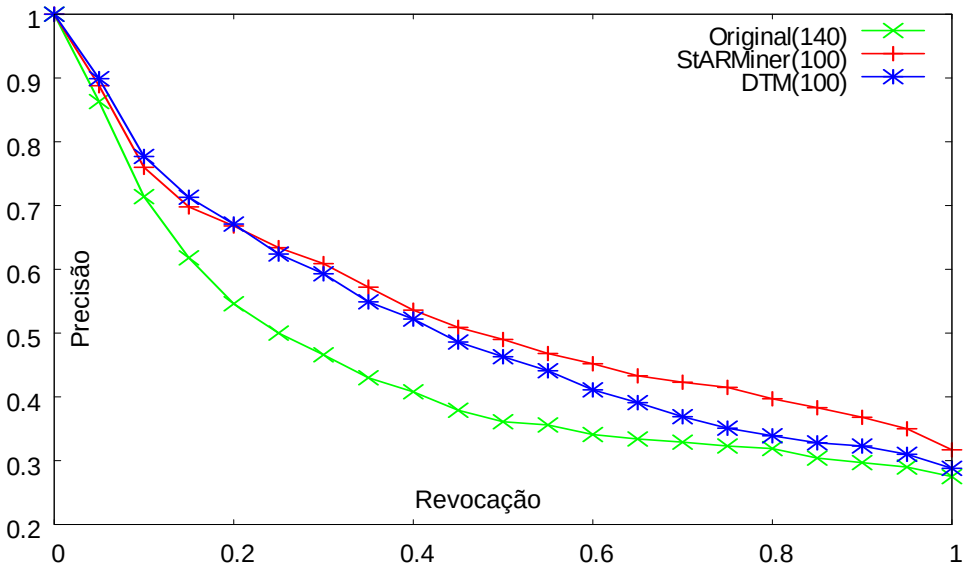


Figura 5.12: Curvas de P&R obtidas utilizando o conjunto TexturaRMI943 representado por: (a) as 140 características originais; (b) as 100 características selecionadas pelo StARMiner; e (c) as 100 selecionadas pelo algoritmo DTM.

O gráfico da figura 5.12 mostra um aumento nos valores de precisão com a redução de dimensionalidade. Este resultado é mais uma evidência de que a mineração de regras de associação pode ser empregada com sucesso para seleção de características, melhorando a qualidade das consultas nos sistemas de CBIR.

5.3.5 Estudo de Caso 4

Neste estudo de caso o algoritmo StARMiner é utilizado para **ponderar** as características para representar a base TexturaRMI704. A base foi dividida em dois conjunto de dados: conjunto de

treino, composto por 25% das imagens, e o conjunto de teste, composto por 75% das imagens. Os parâmetros utilizados pelo StARMiner foram $\Delta\mu_{min} = 0.2$, $\sigma_{max} = 0.1$ e $\gamma_{min} = 0.99$.

Os gráficos de P&R da Figura 5.13 correspondem aos experimentos realizados utilizando a base TexturaRMI704, onde os itens (a), (b), (c), (d), (e) e (f) correspondem aos resultados usando as funções de distância L_1, L_2, L_{inf}, χ^2, Divergência de Jeffrey e Canberra (dicutidas no capítulo 3), comparando a ponderação por regras de associação com vetores não ponderados e com outros algoritmos de seleção de características. As curvas de P&R dos gráficos da figura 5.13 foram construídas executando consultas por similaridade empregando os seguintes critérios de seleção de características:

- vetor original;
- seleção pelo StARMiner (removendo características irrelevantes, mas não ponderando);
- seleção e ponderação pelo StARMiner (ponderando as características e removendo as irrelevantes);
- ponderação pelo StARMiner (ponderando as características e mantendo as irrelevantes);
- seleção pelo algoritmo Relief-F;
- seleção pelo algoritmo DTM.

A seleção de características realizada pelo StARMiner levou a uma redução de 20% do vetor de características. O mesmo número de características selecionadas pelo algoritmo StARMiner (112) foi selecionado das mais relevantes retornadas pelos métodos Relief-F e DTM.

A figura 5.13 mostra os gráficos de P&R obtidos. Analisando os gráficos é possível observar que a ponderação de características pelo StARMiner melhora a precisão das consultas por conteúdo, e promove um aumento de precisão maior do que as técnicas de seleção sem ponderação utilizadas (seleção pelo StARMiner, Relief-F e DTM).

Na figura 5.13 (b), a ponderação pelo StARMiner atinge um ganho considerável, de aproximadamente 20% utilizando ponderação e seleção e de 38% utilizando somente a ponderação, para valores de revocação de 35%. Esses resultados ilustram que a ponderação pelo StARMiner, mesmo reduzindo o tamanho do vetor de características, aumenta a precisão das consultas por similaridade.

Pela análise da figura 5.13, é possível observar que a função de distância que produz os menores valores de precisão é a L_{inf} (figura 5.13 (c)). Mesmo utilizando essa função de distância, a ponderação de características pelo algoritmo StARMiner leva a um considerável ganho na precisão. A função de distância que produz o melhor resultado é a *Canberra* (figura 5.13 (f)). Para a distância *Canberra*, a ponderação de características pelo algoritmo StARMiner também levou a um ganho na precisão.

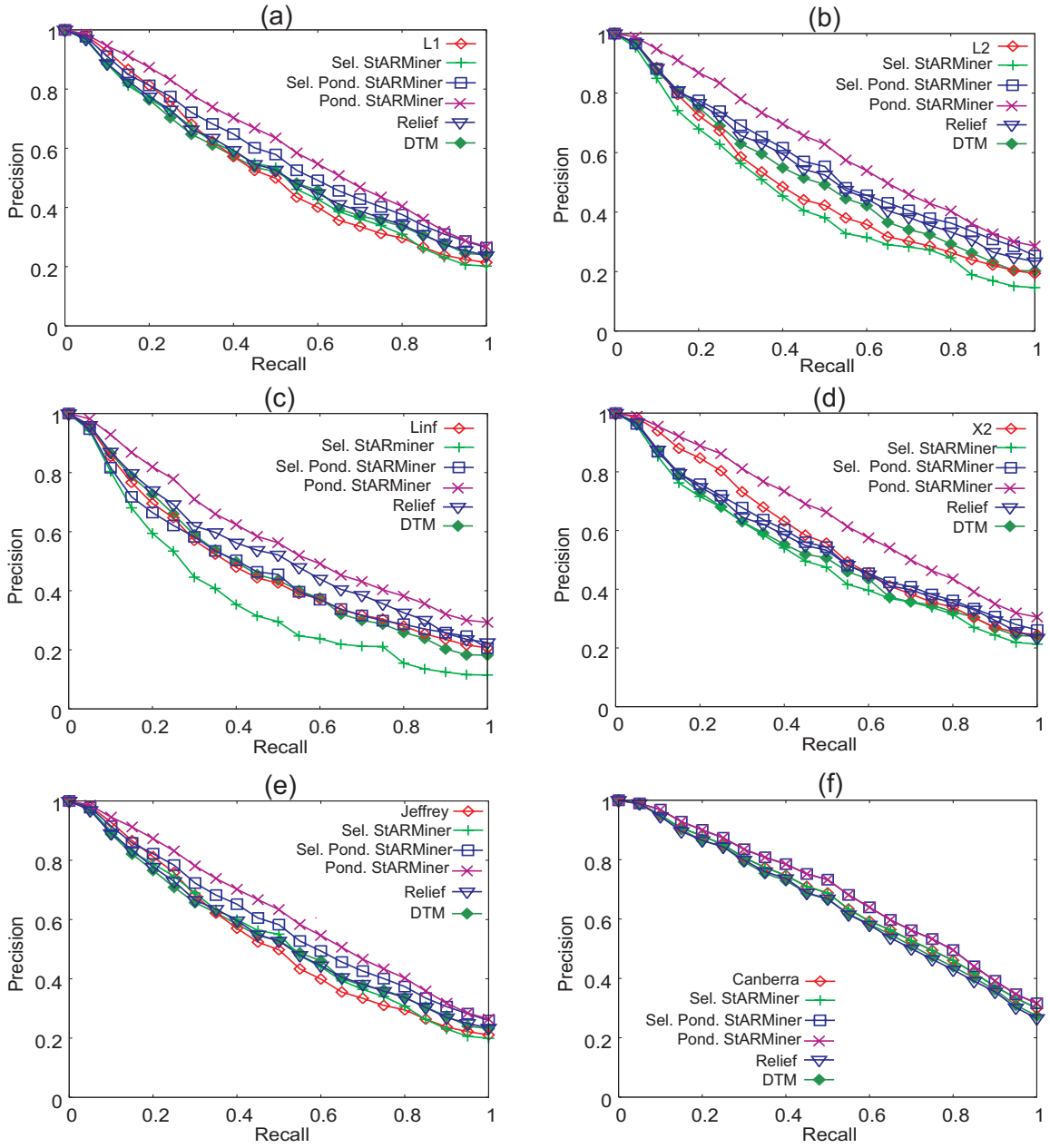


Figura 5.13: Gráficos de P&R usando as funções de distâncias (a) L_1, (b) L_2, (c) L_{inf}, (d) χ^2, (e) Divergência Jeffrey e (f) Canberra obtidas sobre a base TexturaRMI704, empregando os seguintes critérios de seleção de características: vetor original; seleção pelo StARMiner (removendo características irrelevantes, mas não ponderando); seleção e ponderação pelo StARMiner (ponderando as características e removendo as irrelevantes); ponderação pelo StARMiner (ponderando as características e mantendo as irrelevantes); seleção pelo algoritmo Relief-F; e seleção pelo algoritmo DTM.

A figura 5.14 mostra um exemplo de consulta *k*NN (*k*=8), onde a imagem do canto superior esquerdo é utilizada como centro de consulta. A figura 5.14 (a) mostra os resultados obtidos usando os vetores de características originais, e a figura 5.14 (b) mostra os resultados usando ponderação e seleção de características pelo algoritmo StARMiner. As imagens contornadas por uma linha tracejada indicam falsos positivos. Um falso positivo é uma imagem retornada cuja classe se difere da classe centro de consulta. Claramente, a ponderação de características pelo algoritmo StARMiner produz os melhores resultados para essa consulta.

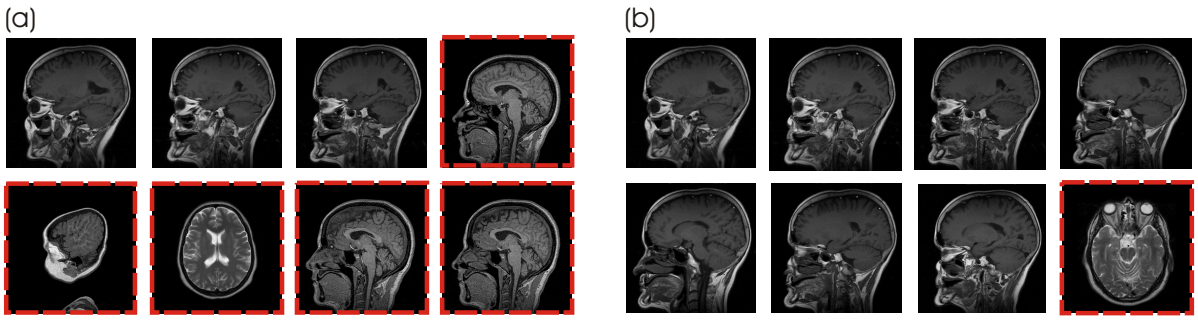


Figura 5.14: Um exemplo de consulta *k*NN (*k*=8) usando a função de distância L_2, onde a imagem do canto superior esquerdo é o centro de consulta. (a) usando o vetor de características original; (b) usando ponderação e seleção de características pelo algoritmo StARMiner. As imagens contornadas por uma linha tracejada são falsos positivos.

É importante notar que, além de utilizar as distâncias tradicionais da família L_p os experimentos também foram executados utilizando as funções de distância χ^2, Divergência de Jeffrey e Canberra (figuras 5.13 (d), (e) and (f)). Esses resultados indicam que a seleção e ponderação de características pelo algoritmo StARMiner pode ser estendida para outras funções de distância, além da tradicional família L_p (família Minkowski), e, embora não apresentados aqui, a outros tipos de vetores de características, apresentando um ganho de precisão notável nas consultas por similaridade.

Este estudo de caso mostra que a seleção e ponderação de características pelo algoritmo StARMiner, quase sempre, produz um maior aumento na precisão do que os algoritmos tradicionais de seleção de características Relief-F e DTM. Considerando os resultados alcançados, é possível afirmar que a mineração de regras de associação é bastante adequada para promover seleção e ponderação de vetores de características, aumentando a precisão das buscas por conteúdo em imagens médicas.

5.4 Considerações Finais

A técnica de mineração aqui apresentada atua na extração de padrões dos vetores de características empregados para comparar e recuperar as imagens por similaridade. Através dos padrões

minerados é possível selecionar e/ou ponderar as características que representam as imagens. Embora a técnica aqui apresentada seja direcionada para minerar imagens médicas, ela pode ser diretamente estendida para trabalhar com outros tipos de dados.

Os estudos de caso apresentados neste capítulo mostram a aplicabilidade da mineração de regras de associação para a redução de dimensionalidade de bases de imagens, além de permitir a mineração de padrões que relacionam características de baixo nível das imagens com informação de alto nível das imagens (categorias). Os padrões gerados pelo algoritmo StARMiner podem auxiliar os pesquisadores da área de processamento de imagens a entender melhor o comportamento das características de baixo nível e a determinar quais delas são realmente importantes para descrever o conteúdo da imagem. As regras mineradas também detalham o comportamento das características nos subconjuntos de imagens. Nos experimentos descritos, a seleção de características realizada pelo algoritmo StARMiner também é comparada com outros algoritmos bastante conhecidos de seleção de características. Os resultados indicam que o algoritmo StARMiner é bastante adequado para a tarefa de seleção de características em bases de imagens médicas.

Outros trabalhos, não apresentados neste capítulo, foram realizados utilizando o algoritmo StARMiner.

No artigo [Ribeiro et al., 2006c], o algoritmo StARMiner é utilizado em conjunto com técnicas de realimentação de relevância para maximizar a precisão das buscas por conteúdo, onde dois problemas inerentes a busca por conteúdo são tratados: a "maldição da alta-dimensionalidade" e o "gap semântico". Esses dois problemas são tratados acoplando a seleção de características promovida pelo algoritmo StARMiner ao processo de realimentação de relevância atingindo resultados promissores.

No artigo [Sousa et al., 2006], a mineração de regras de associação estatísticas é aplicada em *data streams*. Nesse artigo, a dimensão intrínseca (dimensão fractal) do fluxo dos dados é monitorada e, quando existe alguma mudança, o algoritmo StARMiner é acionado para atualizar as regras mineradas sobre o fluxo dos dados. Assim, as regras de associação são mineradas somente quando existe alguma alteração no comportamento dos dados indicada por uma mudança no valor de sua dimensão intrínseca.

Capítulo 6

O Algoritmo Omega

6.1 Considerações Iniciais

Este capítulo apresenta um novo algoritmo que permite pré-processar as características de baixo nível automaticamente extraídas das imagens para a mineração de regras de associação.

O pré-processamento de dados é um elemento chave para aumentar a precisão dos algoritmos de mineração. Na fase de pré-processamento, os dados são tratados de maneira a tornar o processo de mineração possível e mais preciso. A discretização de dados e a seleção de características são duas tarefas importantes que podem ser realizadas no pré-processamento dos dados e podem reduzir significativamente o esforço do algoritmo de mineração.

Neste capítulo é apresentado o algoritmo Omega, um novo algoritmo supervisionado que realiza *discretização* e *seleção de características* de uma maneira bastante eficiente. O objetivo deste capítulo é explicar detalhadamente o algoritmo Omega e mostrar a aplicação do mesmo, de uma maneira genérica para discretização de dados e seleção de características, comparando-o com outros algoritmos tradicionais da literatura. A aplicação do algoritmo Omega no método de auxílio ao diagnóstico desenvolvido nesta tese é apresentada no próximo capítulo.

Aqui, Omega é validado através da comparação com outros algoritmos bastante conhecidos de discretização (1R, ChiMerge e Chi2) e de seleção de características (DTM, Relief-F e Chi2). Os experimentos compararam os efeitos das técnicas de pré-processamento nos resultados do algoritmo C4.5. Nos resultados, a discretização dos dados promovida pelo algoritmo Omega gera a árvore de decisão com o menor número de nós, e a seleção de características promovida por Omega leva a uma das menores taxas de erros. Assim, os resultados dos experimentos mostram que o algoritmo Omega é bastante adequado para realizar seleção de características e discretização de dados, sendo apropriado para pré-processar os dados para as tarefas de mineração de dados.

6.2 Discretização e Seleção de Características

Os tipos mais comuns de características (atributos) usados na mineração de dados são: as nominais, as contínuas e as discretas. As características nominais freqüentemente assumem um número limitado de valores, porém não apresentam relação de ordem entre eles. As características contínuas podem assumir um número infinito de valores, porém preservam a relação de ordem entre eles. Para muitos algoritmos de mineração é importante que as características atendam ambos os requisitos: um número limitado de valores e a preservação de ordem entre os valores, aumentando a velocidade e a qualidade do processo de mineração. Esses dois requisitos estão presentes em dados discretos, cujo valores preservam a relação de ordem e freqüentemente tem um número limitado de valores. Assim, um importante passo do pré-processamento dos algoritmos de mineração é o processo de discretização de valores contínuos.

Outro passo importante do pré-processamento é a seleção de características. A seleção de características elimina características irrelevantes e redundantes que interferem negativamente no resultado da mineração. Diferente do que é esperado pelo senso comum, a utilização de um grande número de características para representar um objeto (uma imagem, por exemplo) é um problema. Com o aumento do número de características, a significância de cada característica diminui e o tempo gasto para a análise dos dados aumenta, levando a ocorrência dos efeitos da "maldição da alta dimensionalidade". Além disso, em muitos casos, um grande número de características são correlacionadas, carregando informação redundante que ao invés de auxiliar, atrapalha o processo de mineração de dados.

O capítulo 4 apresenta uma revisão sobre as técnicas de discretização e seleção de características existentes na literatura. Dos algoritmos discutidos no capítulo 4, somente o Chi2 realiza simultaneamente seleção de características e discretização. Os demais algoritmos discutidos realizam somente seleção de características ou somente discretização, requerendo, na maioria das vezes, um algoritmo adicional para completar a operação de pré-processamento sobre os dados.

Neste capítulo é apresentado o algoritmo Omega que usa uma medida de inconsistência para determinar o número final de intervalos e para selecionar as características. Quando o número de intervalos diminui, o número de inconsistência aumenta. Omega tenta manter o número mínimo de intervalos com a menor taxa de inconsistência, estabelecendo uma relação de compensação entre essas duas medidas. Uma vantagem do algoritmo Omega sobre seus precedentes é seu baixo custo computacional. Omega tem um custo linear para promover seleção de características e discretização de um conjunto de N valores ordenados. Como, o Chi2 também realiza ambas as tarefas executadas pelo algoritmo Omega, a comparação entre os algoritmos Omega e Chi2 é focada nos experimentos realizados.

6.3 Descrição do Algoritmo Omega

Omega é um novo algoritmo supervisionado que realiza discretização e seleção de características. O algoritmo Omega processa cada característica separadamente. A seguir, os passos do algoritmo Omega são apresentados. Os passos 1 até 3 são empregados para realizar discretização de dados, enquanto o passo 4 realiza seleção de característica.

Seja f uma característica e f_i o valor da característica f em uma tupla i. Omega usa uma estrutura de dados que associa o valor de f_i em cada tupla com a classe da instância c_i. De agora em diante, o termo instância I_i passa a referenciar o par (f_i, c_i). Seja U_k e U_{k+1} os limites de um intervalo T_k.

Definição 6.1 *Uma instância $I_i = (f_i, c_i)$ pertence a um intervalo $T_k = [U_k, U_{k+1}]$ se e somente se $U_k < f_i < U_{k+1}$.*

O algoritmo Omega requer que os valores das características estejam previamente ordenados.

Passo 1

No Passo 1, o Omega define os pontos de corte (limites dos intervalos) iniciais. Primeiramente, um ponto de corte é colocado antes do menor valor e outro ponto de corte é colocado depois do maior valor da característica. Em seguida, toda vez que o valor da característica alterar e ocorrer uma mudança de classe, um ponto de corte é criado.

No Passo 1, Omega produz intervalos (*bins*) puros, que são intervalos onde a entropia é a menor possível (zero). O Passo 1 produz intervalos que minimiza a inconsistência adquirida no processo de discretização. Entretanto, o número de intervalos gerados neste primeiro passo tende a ser bastante grande e também bastante sujeito a ruídos.

Um processo de discretização que produz um grande número de intervalos (no pior caso, o mesmo número de valores contínuos original) não é desejável, pois não adiciona nenhum ganho ao algoritmo de mineração. Assim, depois do Passo 1, Omega passa a eliminar pontos de corte de maneira a reduzir o número de intervalos, controlando fortemente a taxa de inconsistência, mantendo-a abaixo de um determinado limiar.

A figura 6.1 mostra um exemplo dos valores originais de uma característica, ordenados de maneira ascendente, e os pontos de corte encontrados no Passo 1 do algoritmo Omega. As linhas tracejadas representam pontos de corte.

Passo 2

No Passo 2, Omega restringe a freqüência mínima que um intervalo deve satisfazer. Essa restrição funciona como um filtro inicial, evitando a criação de um grande número de pontos de

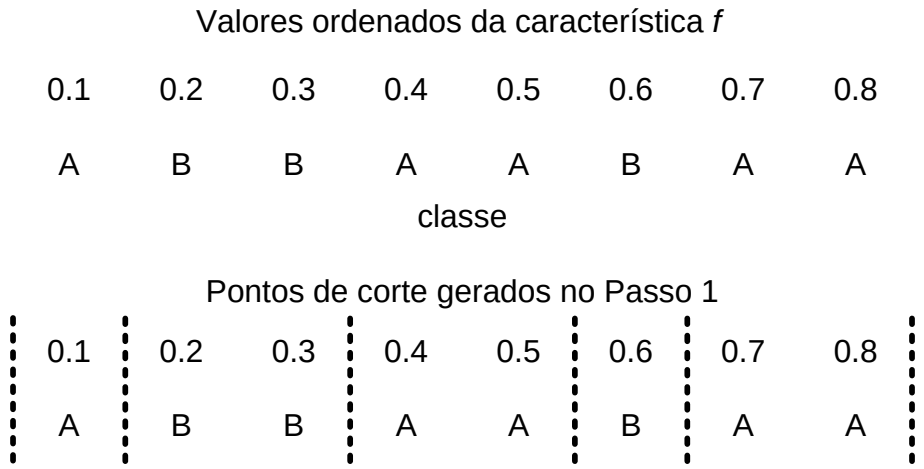


Figura 6.1: Exemplo de pontos de corte criados no Passo 1 do algoritmo Omega.

corte. Omega remove o ponto de corte à direita dos intervalos que não satisfazem a restrição de freqüência mínima fornecida por um parâmetro de entrada H_{min}. Apenas é permitido ao último intervalo não satisfazer a restrição de freqüência mínima.

Quanto maior o valor de H_{min}, menor o número de intervalos obtidos neste passo. Entretanto, o valor desse parâmetro deve ser ajustado com cautela, porque quanto maior o valor de H_{min}, maior a quantidade de inconsistência gerada pelo processo de discretização. Assim, é importante que o valor de H_{min} seja mantido baixo (porém maior que 1), mesmo que seja atingida apenas uma pequena redução no número de intervalos. O próximo passo do algoritmo garante uma maior redução no número de intervalos, controlando o número de inconsistências que é gerada pelo processo de discretização. A figura 6.2 mostra um exemplo dos pontos de corte encontrados no Passo 1 (ilustrados na figura 6.1) que são eliminados no Passo 2 do algoritmo Omega, usando $H_{min} = 2$.

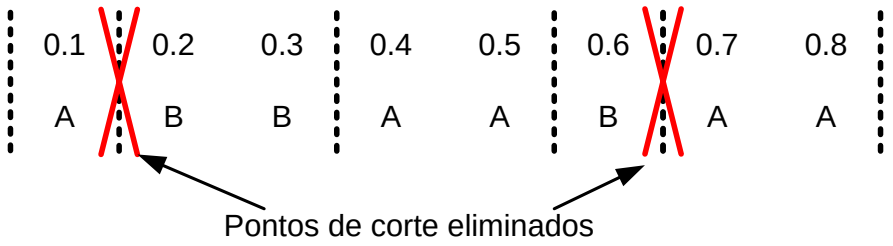


Figura 6.2: Pontos de corte eliminados no Passo 2 do algoritmo Omega, usando $H_{min} = 2$.

Passo 3

No Passo 3, o algoritmo Omega junta intervalos consecutivos, limitando a taxa de inconsistência para realizar a fusão. Diferente do algoritmo Chi2, que usa o teste estatístico χ^2

para determinar quando unir intervalos, Omega usa a medida de taxa de inconsistência para determinar quais intervalos devem ser unidos. Seja M_{T_k} a classe majoritária de um intervalo T_k. A classe majoritária é a classe mais freqüente de um intervalo. A equação 6.1 fornece a taxa de inconsistência ζ_{T_k} para um intervalo T_k.

$$\zeta_{T_k} = \frac{|T_k| - |M_{T_k}|}{|T_k|} \tag{6.1}$$

Na equação 6.1, $|T_k|$ é o número de instâncias no intervalo T_k e $|M_{T_k}|$ é o número de instâncias da classe majoritária no intervalo T_k. O algoritmo Omega une intervalos consecutivos que possuem a mesma classe majoritária e que possuem uma taxa de inconsistência menor ou igual a um limiar de entrada ζ_{max} ($0 \leq \zeta_{max} \leq 0.5$). A figura 6.3 mostra um exemplo de ponto de corte encontrado no Passo 2 (veja figura 6.2) que é eliminado no Passo 3 do algoritmo Omega, usando $\zeta_{max} = 0.35$. As taxas de inconsistência ζ_{T_k} do segundo e do terceiro intervalos mostrados na figura 6.3 são respectivamente $\zeta_{T_2} = 0/2 = 0$ e $\zeta_{T_3} = 1/3 = 0.33$. Desde que T_2 e T_3 têm a mesma classe majoritária, isto é, $M_{T_2} = M_{T_3}$ =“A” e $\zeta_{T_2} \leq \zeta_{max}$ e $\zeta_{T_3} \leq \zeta_{max}$, o segundo e o terceiro intervalos são unidos. Os pontos de corte restantes no Passo 3 são os pontos de corte finais retornados pelo algoritmo.

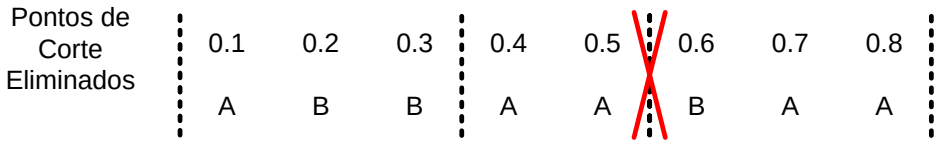


Figura 6.3: Um ponto de corte eliminado no Passo 3 do algoritmo Omega, usando $\zeta_{max} = 0.35$.

Passo 4 (Seleção de Características)

No **Passo 4**, o algoritmo Omega realiza seleção de características. Seja T o conjunto de intervalos que uma característica é discretizada. Para cada característica, Omega calcula o valor da inconsistência global ζ_G, de acordo com a equação 6.2.

$$\zeta_G = \frac{\sum_{T_k \in T}(|T_k| - |M_{T_k}|)}{\sum_{T_k \in T}|T_k|} \tag{6.2}$$

O **critério de seleção de características** empregado por Omega remove todas as características cujo valor de inconsistência global é maior do que um limiar de entrada $\zeta_{G_{max}}$ ($0 \leq \zeta_{G_{max}} \leq 0.5$). De fato, a quantidade de inconsistência é o fator que mais contribui para

distorcer os resultados de um algoritmo de mineração. Assim, o descarte dos atributos mais inconsistentes pode contribuir para aumentar a precisão e a velocidade dos algoritmos de mineração.

A figura 6.4 mostra os pontos de corte finais determinados pelo algoritmo Omega ao processar os valores de uma característica f apresentados na figura 6.1. Como, de oito instâncias, apenas duas têm classes diferentes da classe majoritária de seus intervalos, a inconsistência global da característica f (ilustrada na figura 6.4) é $\zeta_G = 2/8 = 0.25$. Se $\zeta_G \leq \zeta_{G_{max}}$, a característica é selecionada, caso contrário, ela é eliminada do vetor de características. Assim, se $\zeta_{G_{max}} \geq 0.25$ então f é selecionada, caso contrário, f é descartada do vetor de características.

Pontos de Corte Finais	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	A	B	B	A	A	B	A	A

Figura 6.4: Pontos de corte finais encontrados pelo algoritmo Omega.

Omega é apresentado no algoritmo 3. É importante ressaltar que Omega discretiza N valores ordenados em $4N$ passos.

Algoritmo 3: Algoritmo Omega.

Dados: Conjunto de N instâncias $I_i = (f_i, c_i)$ ordenadas por f_i, parâmetros H_{min}, ζ_{max} e $\zeta_{G_{max}}$.

Resultado: Conjunto U de pontos de corte encontrados, *selected* (**false**, se a característica deve ser eliminada, e **true**, caso contrário).

1 (**Passo 1**) Adicione a U pontos de corte U_k antes de f_0, depois de f_{N-1} e entre todos f_i e f_{i+1} **para** $f_i \neq f_{i+1}$ **e** $c_i \neq c_{i+1}$;
2 (**Passo 2**) **Se** U_{k+1} não é o último intervalo **então remova** de U os pontos de corte U_{k+1} de todo o intervalo $T_k = [U_k, U_{k+1}]$ que possui menos que H_{min} instâncias;
3 (**Passo 3**) Remova de U o ponto de corte U_{k+1} entre todos os intervalos consecutivos $T_k = [U_k, U_{k+1}]$ **e** $T_{k+1} = [U_{k+1}, U_{k+2}]$ **se** $M_{T_k} = M_{T_{k+1}}$ **e** $\zeta_{T_k} \leq \zeta_{max}$ **e** $\zeta_{T_{k+1}} \leq \zeta_{max}$;
4 (**Passo 4**) Calcule ζ_G;
5 **se** $\zeta_G \leq \zeta_{G_{max}}$ **então**
6 | *selected*=true;
7 **fim**
8 **senão**
9 | *selected*=false;
10 **fim**
11 **retorna** *U e selected*

6.4 Experimentos

Foram realizados vários experimentos para validar o algoritmo Omega. Nos experimentos apresentados aqui foram utilizados os conjuntos de dados *Australian* (690 tuplas, 13 características, 2 classes), *Heart* (270 tuplas, 14 características, 2 classes) e *Vehicle* (846 tuplas, 18 características, 4 classes). Esses conjuntos de dados são públicos e estão disponíveis no repositório da UCI [Asuncion & Newman, 2007].

Foram comparados os efeitos de submeter os dados pré-processados por Omega e outros algoritmos conhecidos de pré-processamento ao classificador C4.5. O algoritmo C4.5 foi escolhido para validar Omega porque o C4.5 pode lidar com ambos os tipos de atributos, contínuos e discretos, e porque o C4.5 se tornou um padrão para comparações no campo da mineração de dados. No primeiro estudo de caso, o algoritmo Omega foi validado na tarefa de discretização de dados, comparando ele com outros métodos de discretização. No segundo estudo de caso, o algoritmo Omega foi validado na tarefa de seleção de características, comparando-o com outros métodos de seleção.

6.4.1 Estudo de Caso 1

Neste estudo de caso, o algoritmo Omega é validado na tarefa de discretização de dados. Para realizar essa validação, o número de nós gerados na árvore de decisão são comparados, usando os dados não pré-processados e quando os dados são pré-processados utilizando um método de discretização.

Os conjuntos de dados foram pré-processados utilizando cinco diferentes métodos de discretização: *equal-sized* (os dados contínuos foram discretizados em 10 intervalos), 1R, ChiMerge, Chi2 e Omega. Os parâmetros dos algoritmos 1R, ChiMerge, Chi2 e Omega foram calibrados para maximizar a precisão alcançada pelo algoritmo C4.5. A implementação dos métodos 1R, ChiMerge e Chi2 foi obtida de `www.public.asu.edu/~huanliu/`. Para discretizar um conjunto de N valores ordenados, os algoritmos Omega e 1R tem complexidade $O(N)$ e os algoritmos ChiMerge e Chi2 têm complexidade $O(NlogN)$.

A tabela 6.1 mostra os valores de média e desvio padrão das taxas de erro do algoritmo C4.5 executado usando a validação *k-fold* (k=10) (a) sem utilizar pré-processamento (linha Nada) e utilizando os métodos de pré-processamento: (b) *equal-sized*, (c) 1R, (d) ChiMerge, (d) Chi2 e (f) Omega. A tabela 6.2 mostra a média do número de nós na árvore de decisão obtida usando os mesmos métodos de discretização. A figura 6.5 mostra uma comparação gráfica entre os valores da taxa de erros e o número de nós da árvore de decisão gerados pelo algoritmo C4.5 utilizando os métodos de discretização.

Um processo de discretização deve adicionar o mínimo possível de inconsistência aos dados,

Tabela 6.1: Média da taxa de erro (%) e desvio padrão do algoritmo C4.5 sem utilizar um método de discretização (linha Nada) e utilizando os métodos de discretização: *equal-sized*, 1R, ChiMerge, Chi2 e Omega.

Método	**Média da taxa de erros (%) e desvio padrão**
Nada	21.5 ± 4.8
Equal-sized(k=10)	23.3 ± 6.4
1R	20.5 ± 6.4
ChiMerge	21.9 ± 6.8
Chi2	22.3 ± 8.3
Omega	$\mathbf{21.3 \pm 5.6}$

Tabela 6.2: Média do número de nós na árvore de decisão gerada pelo algoritmo C4.5 sem utilizar um método de discretização (linha Nada) e utilizando os métodos de discretização: *equal-sized*, 1R, ChiMerge, Chi2 e Omega.

Método	**Média do número de nós**
Nada	97.3
Equal-sized(k=10)	184.3
1R	90.3
ChiMerge	95
Chi2	75.3
Omega	**57.3**

visando não aumentar muito a taxa de erro. Além disso, quanto menor o número de nós na árvore de decisão, mais rápida é a fase de aprendizado e de teste do algoritmo de classificação. Assim, é esperado que o melhor método de discretização produza ambos, a menor taxa de erros e o menor número de nós na árvore de decisão. Entretanto, os resultados apresentados nas tabelas 6.1 e 6.2 mostram que nenhum método pode produzir ambos, a menor taxa de erros e o menor número de nós. A tabela 6.1 indica que o método de discretização que produz a menor taxa de erros é o algoritmo 1R, enquanto a tabela 6.2 indica que o algoritmo que produz o menor número de nós na árvore de decisão é o algoritmo Omega. Omega produz uma média de taxa de erro 0.8% maior que a taxa de erro do 1R. Entretanto, o algoritmo 1R leva a um aumento de aproximadamente 57% sobre o número de nós que o algoritmo Omega produz (observe a figura 6.5).

O pior método de discretização é o método *equal-sized*, pois ele apresenta a maior taxa de erros e também gera a maior árvore de decisão (ver tabelas 6.1 e 6.2). Este fato mostra que o uso de um método de discretização não adequado pode aumentar em muito a taxa de erros e a complexidade do modelo de aprendizado gerado pelo classificador.

Muito embora, os algoritmos Omega e Chi2 também realizem seleção de características, estes algoritmos foram empregados sem realizar essa tarefa neste estudo de caso. O número de nós obtidos usando Omega e Chi2 foram os menores quando comparados com os outros méto-

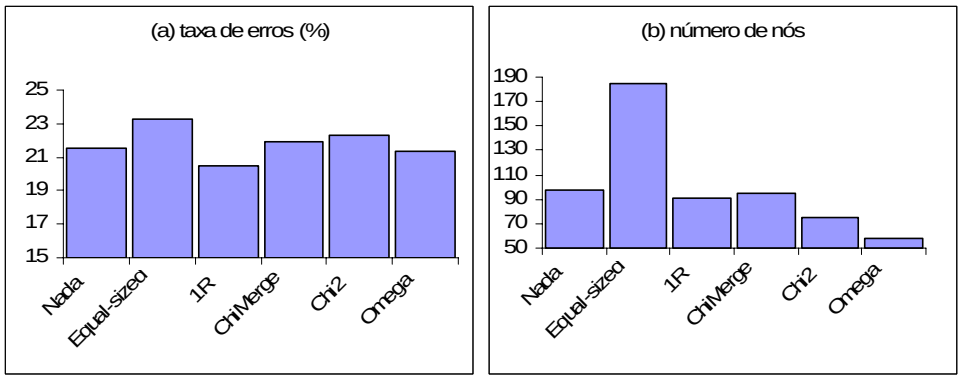


Figura 6.5: Comparação entre as taxas de erro (a) e o número de nós na árvore de decisão (b) gerados pelo algoritmo C4.5 sem utilizar um método de discretização (coluna Nada) e utilizando os métodos de discretização: *equal-sized*, 1R, ChiMerge, Chi2 e Omega.

dos (ver tabela 6.2). Este fato ocorre porque Omega e Chi2 permitem uma quantia controlada de inconsistência para ser adicionada nos dados, tendo uma maior autonomia para reduzir o número de intervalos obtidos durante o processo de discretização. Usando menos intervalos, com a taxa de inconsistência controlada, é esperado que a árvore de decisão construída seja menor. Entretanto, ao comparar Omega e Chi2 (ver as tabelas 6.1 e 6.2), Omega produz a menor taxa de erros (1% menor que Chi2) e o menor número de nós na árvore de decisão (24% menos nós do que Chi2).

6.4.2 Estudo de Caso 2

Neste estudo de caso, o algoritmo Omega foi validado na tarefa de seleção de características. Para realizar essa validação, foram comparadas as taxas de erro obtidas pelo algoritmo C4.5 ao utilizar os dados de entrada sem pré-processar e quando os dados de entrada são pré-processados usando quatro diferentes métodos de seleção de características: Relief-F, DTM, Chi2 e Omega.

Neste estudo de caso, Omega e Chi2 foram empregados apenas para a realização de seleção de características. O algoritmo Omega foi executado sobre o conjunto de dados sendo os parâmetros calibrados para minimizar a taxa de erros ao executar o algoritmo C4.5, selecionando um dado número de características. Os parâmetros do algoritmo Relief-F, DTM e Chi2 foram ajustados de maneira a selecionar o mesmo número de características que o algoritmo Omega selecionou. O número de características *eliminadas* para cada base de dados foram: 10 para a base *Australian*, 8 para a base *Heart* e 5 para a base *Vehicle*.

A tabela 6.3 mostra as médias dos valores de taxa de erro e desvio padrão obtidas pelo algoritmo C4.5 (empregando validação *k-fold*, com k=10) quando utilizando como entrada apenas o conjunto de características selecionadas em cada método para representar cada base de dados. A figura 6.6 mostra uma comparação gráfica entre as médias das taxas de erro atingidas pelos

métodos de seleção de características.

Tabela 6.3: Média das taxas de erro (%) e desvio padrão do algoritmo C4.5 obtidas sem utilizar seleção de característica (linha Nada) e utilizando os métodos de seleção de características: Omega, Chi2, Relief-F e DTM.

Método	**Taxa de erro média e desvio padrão**
Nada	21.5 ± 4.8
Chi2	20.7 ± 6.03
Relief-F	21.2 ± 6.34
DTM	21.0 ± 5.52
Omega	$\mathbf{20.2} \pm 6.12$

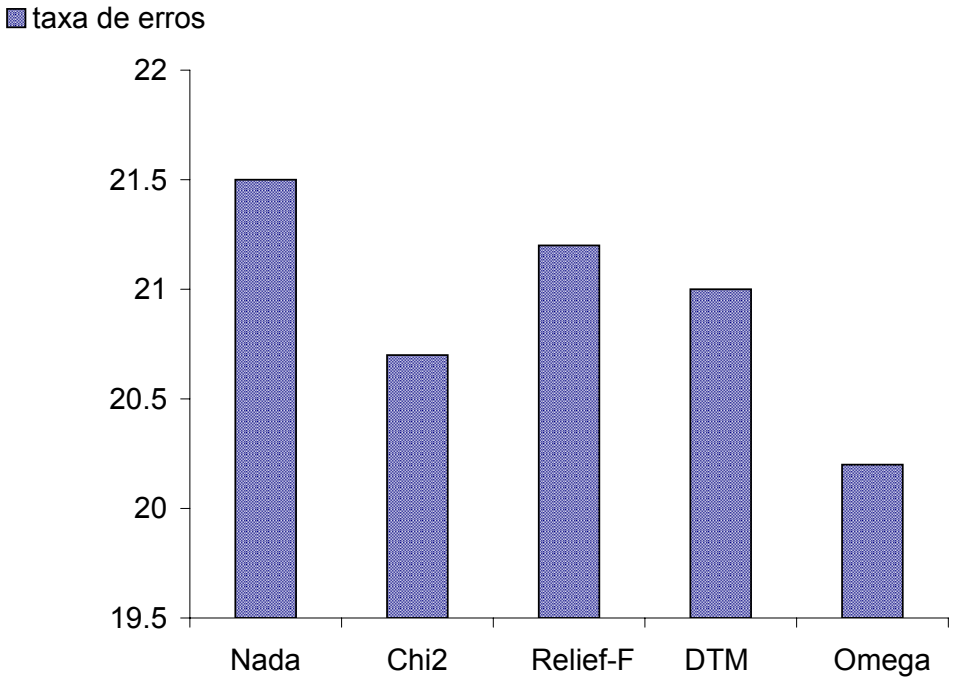


Figura 6.6: Comparação entre a taxa de erro média alcançada pelos métodos de seleção de características.

Os resultados da tabela 6.3 indicam que todos os algoritmos de seleção de características testados reduzem a taxa de erros do algoritmo C4.5. Esse resultado evidencia que a "maldição da alta dimensionalidade" realmente prejudica os resultados do algoritmo de mineração: os atributos irrelevantes distorcem a tendência dos atributos relevantes. Note que, embora os algoritmos Omega e Chi2 também promovem discretização, eles foram utilizados sem realizar esta tarefa neste estudo de caso. Entre os algoritmos testados, Omega alcança os melhores resultados, produzindo a menor taxa de erros (ver figura 6.6).

Entre os algoritmos testados, apenas Chi2 e Omega realizam as tarefas de discretização e seleção de características. Comparando Omega com Chi2 em ambos os estudos de caso 1 e

2, Omega alcança os melhores resultados: a menor taxa de erros e o menor número de nós na árvore de decisão (ver tabelas 6.1, 6.2, e 6.3). Também, a complexidade do algoritmo Omega é menor que a complexidade do algoritmo Chi2. Esses resultados indicam que o algoritmo Omega é bastante adequado para realizar ambas as tarefas, seleção de características e discretização, sendo apropriado para pré-processar os dados para os algoritmos de mineração.

6.5 Um breve histórico

Durante o trabalho desta tese, antes do algoritmo Omega, foi desenvolvido o algoritmo PreSAGe. O algoritmo PreSAGe foi uma tentativa inicial de desenvolver um algoritmo para o pré-processamento das características automaticamente extraídas das imagens médicas para a tarefa de mineração de regras de associação. O algoritmo Omega surgiu de um aprimoramento do PreSAGe, principalmente em relação ao critério de seleção de características:

- No Omega: as características mais inconsistentes são descartadas;
- No PreSAGe: as características com os maiores números de intervalos discretizados são descartadas.

O algoritmo PreSAGe é apresentado no Apêndice B desta tese. No Apêndice B, PreSAGe é validado na tarefa de seleção de características e comparado com outros algoritmos conhecidos da literatura atingindo um bom resultado. No entanto, durante o decorrer do desenvolvimento deste trabalho de pesquisa percebeu-se que seria necessário uma melhor formalização e mudança no critério de seleção de características do algoritmo PreSAGe para que ele melhor executasse as tarefas discretização e seleção de características. Foi assim que surgiu o Omega, e a partir do momento que o Omega foi implementado, ele passou a substituir o PreSAGe nos experimentos realizados e no método de auxílio ao diagnóstico proposto.

6.6 Considerações Finais

Neste capítulo o algoritmo Omega foi apresentado. Omega é um novo algoritmo supervisionado que promove discretização dos dados e seleção de características. Omega utiliza a taxa de inconsistência para determinar quando eliminar os pontos de corte. Uma medida de inconsistência global é utilizada para determinar quais características (atributos) devem ser removidos da base de dados. Os experimentos compararam os efeitos das técnicas de pré-processamento nos resultados do algoritmo C4.5. Os resultados indicam que Omega gera a menor árvore de decisão e também uma das menores taxas de erro. Além disso, Omega é eficiente, tendo um custo linear para discretizar N valores ordenados. Assim, pode-se afirmar que Omega é adequado para

a seleção de características e para a discretização, sendo apropriado para o pré-processamento dos dados para as tarefas de mineração.

No artigo [Romani et al., 2008], o algoritmo Omega é utilizado para discretizar dados climáticos para a tarefa de mineração de regras de associação. Nesse artigo, a teoria fractal é utilizada para seleção de características (através da descoberta de grupos de correlação) e o algoritmo Omega para a discretização. O algoritmo Apriori é executado sobre os dados pré-processados, identificando regras de associação interessantes, que relacionam faixas de temperatura, índice de vegetação e produtividade.

O algoritmo Omega surgiu na tentativa de pré-processar os dados das características de imagens médicas para a tarefa de associação. Neste capítulo, mostramos os resultados da aplicação do algoritmo Omega em base de dados tradicionais, mas no próximo capítulo ele é apresentado como parte do método proposto de auxílio ao diagnóstico de imagens médicas.

Capítulo 7

Método IDEA

7.1 Considerações Iniciais

O principal objetivo desta tese é mostrar que pode-se utilizar com sucesso a mineração de regras de associação para o auxílio ao diagnóstico de imagens médicas. Neste capítulo é descrito o método IDEA (***I**mage **D**iagnosis **E**nhancement through **A**ssociation rules*), um método de auxílio ao diagnóstico de imagens médicas baseado em regras de associação desenvolvido durante este trabalho de pesquisa.

IDEA é um novo método baseado em regras de associação que sugere uma segunda opinião ao radiologista ou um diagnóstico preliminar de uma nova imagem. A segunda opinião pode ser empregada para acelerar o processo de diagnóstico ou para reforçar uma hipótese, aumentando a chance de um tratamento prescrito surtir efeito positivo.

O método IDEA tem a vantagem de promover seleção de características e discretização em um único passo, reduzindo a complexidade da tarefa de mineração. Além disso, o método sugere um conjunto de palavras-chave para compor o diagnóstico de uma dada imagem e utiliza uma medida de interesse (*convicção*) que indica o grau de certeza (probabilidade) de que uma palavra-chave pertence ao diagnóstico final dado pelo radiologista. Além disso, um protótipo foi desenvolvido para incorporar o método IDEA (o Sistema IDEA). Esse protótipo foi utilizado em uma experiência de uso com dois radiologistas, que demonstraram um grande interesse em empregar o sistema no seu dia a dia.

Para dar suporte ao método IDEA dois novos algoritmos foram desenvolvidos: o algoritmo Omega, discutido no capítulo anterior, e o algoritmo ACE, que é discutido neste capítulo. Vários experimentos foram realizados para validar o método proposto. Os resultados indicam que a mineração de regras de associação pode ser empregada com sucesso para aumentar a precisão dos sistemas de auxílio ao diagnóstico (CAD), constituindo uma ferramenta poderosa para suportar o auxílio ao diagnóstico em sistemas médicos.

7.2 Descrição do Método IDEA

O IDEA é um método supervisionado que minera regras de associação relacionando características visuais automaticamente extraídas das imagens com os laudos das imagens de treino. Os laudos em questão são compostos por conjuntos de palavras-chave que descrevem a análise efetuada pelos radiologistas a respeito das imagens. A figura 7.1 mostra o *pipeline* do método IDEA e o algoritmo 4 resume os passos do método.

Algoritmo 4: Os passos do método IDEA.

Dados: Imagens de treino, imagens de teste
Resultado: Laudo (um conjunto de palavras-chave)
1 Extrair as características das imagens de treino;
2 Executar o algoritmo Omega;
3 Minerar regras de associação;
4 Extrair as características das imagens de teste;
5 Executar o algoritmo ACE;
6 Retornar o conjunto de palavras-chave do diagnóstico sugerido;

Na fase de treino (ver algoritmo 4), as características visuais são extraídas das imagens, e vetores de características são usados para representar as imagens (linha 1). Os vetores de características e a *classe* das imagens de treino são submetidos ao algoritmo Omega. A *classe* de uma imagem é a sua mais importante palavra-chave, que é escolhida por um especialista. O algoritmo Omega remove as características irrelevantes do vetor de características e também discretiza as características restantes (linha 2).

Na fase de treino, as palavras-chave do diagnóstico das imagens de treino são adicionados aos seus respectivos vetores de características (processados pelo algoritmo Omega), produzindo uma representação da imagem por registro. Os registros de todas as imagens de treino são submetidas ao algoritmo Apriori [Agrawal & Srikant, 1994] para a mineração de regras de associação (linha 3), limitando a confiança mínima a valores altos. Na fase de teste (linhas 4-6), o vetor de característica da imagem de teste é extraído (linha 4) e submetido ao algoritmo ACE (linha 5), que usa as regras de associação para sugerir possíveis palavras-chave para compor o diagnóstico da imagem de teste. Cada passo do método IDEA é detalhado a seguir.

7.2.1 Extração de características

O primeiro estágio dos sistemas CAD requer a extração das principais características para representar as imagens considerando um critério específico. Uma questão que deve ser respondida ao se trabalhar com imagens médicas é: “quais são as características visuais que melhor

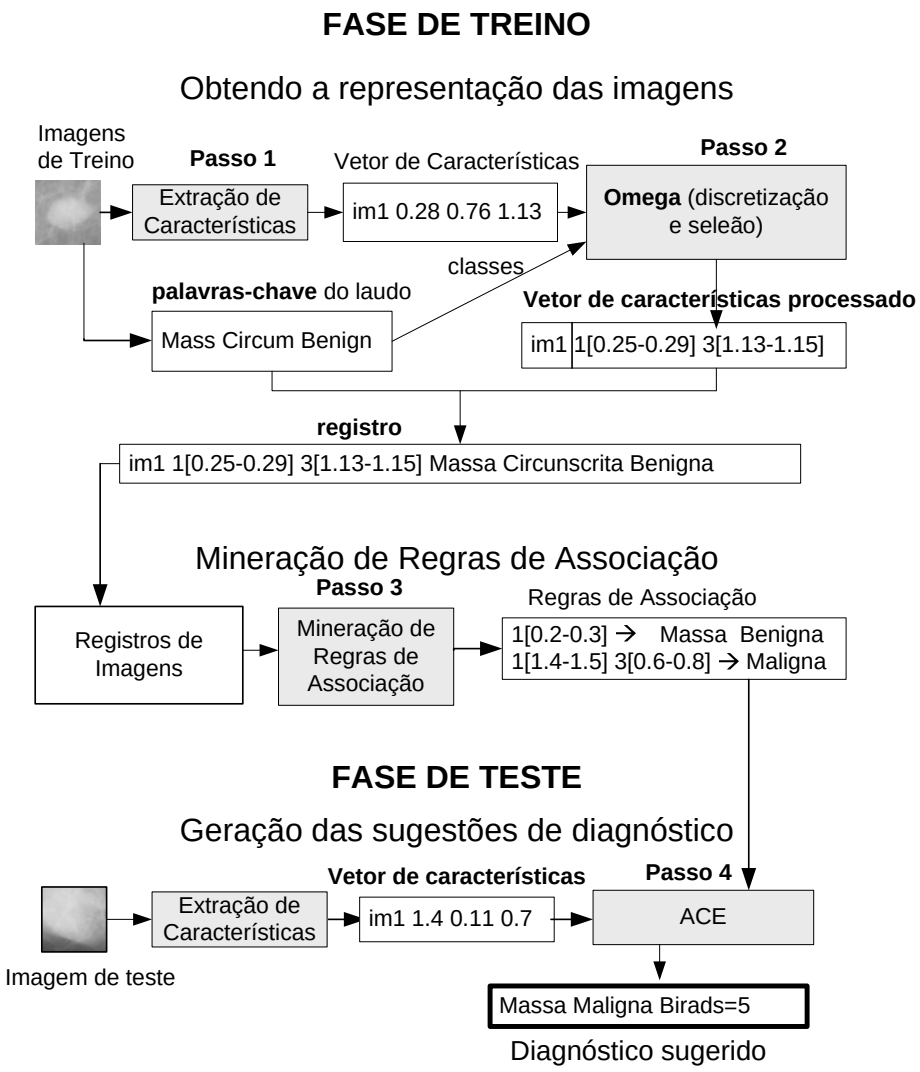


Figura 7.1: *Pipeline* do método IDEA.

representam o significado semântico da imagem?" Essencialmente, as características mais representativas variam com o tipo de imagem (por exemplo, mamografia, ressonância de pulmão e de cérebro) e de acordo com o foco de análise (por exemplo, para distinguir nódulos e para identificar lesões na mielina). Assim, ao se trabalhar com análise de mamografias, as características de cor extraídas do histograma da imagem fornecem uma descrição bastante rudimentar da imagem. Já, as características de forma podem ser adequadas para distinguir lesões malignas e benignas. No entanto, para distinguir a densidade dos tecidos, as características de textura são mais adequadas. Além disso, a maioria das características de forma requer um passo anterior de segmentação das lesões. Esse passo, aumenta bastante o custo computacional e muitas vezes demanda a correção, ou mesmo a delimitação manual da lesão segmentada.

O método IDEA foi desenvolvido para trabalhar com vários tipos de imagens médicas e

com diferentes focos de análise. Entretanto, para cada tipo de imagem e para cada objetivo de análise, um extrator de características apropriado deve ser utilizado. Assim, para cada estudo de caso conduzido para avaliar o método IDEA, um vetor de características apropriado foi utilizado. Os conjuntos de imagens utilizados nos estudos de caso são mamografias e os vetores de característica mais apropriados são descritos no próprios estudos de caso (seção 7.3).

7.2.2 O Algoritmo Omega

O algoritmo Omega já foi descrito no capítulo anterior. A figura 7.2 mostra um exemplo do funcionamento do algoritmo Omega dentro do método IDEA.

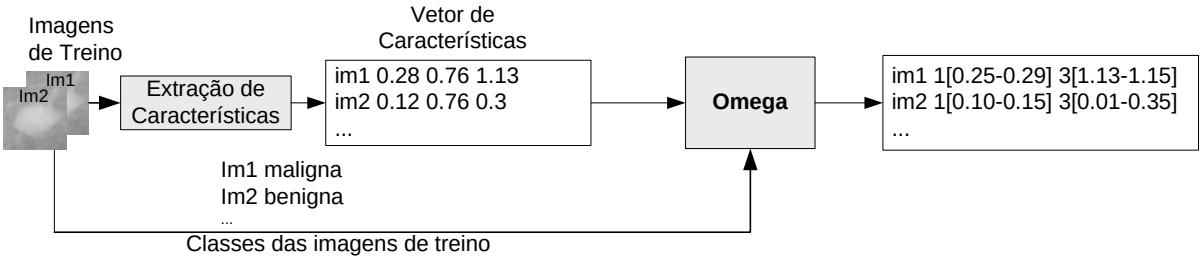


Figura 7.2: Exemplo de funcionamento do algoritmo Omega dentro do método IDEA.

No método IDEA, o algoritmo Omega é empregado para discretizar e selecionar as características das imagens de treino. Para isso os vetores de características das imagens de treino, junto com a classe das imagens são submetidos ao algoritmo. A saída do algoritmo Omega são os vetores de características discretizados. Observe na figura 7.2 que o vetor de características discretizado é composto pelo nome da imagem, a posição da característica discretizada e o intervalo de discretização. Assim, a primeira característica da $Im1$ foi discretizada para o intervalo fechado [0.25-0.29].

Omega também remove as características menos relevantes (aquelas que possuem uma maior taxa de inconsistência) do vetor de características. Um exemplo fictício, mas que mostra a intuição da seleção de características feita pelo algoritmo Omega é apresentado na figura 7.2. Observe que a segunda característica tem o mesmo valor (0.76) nos vetores de características das imagens $Im1$ (maligna) e $Im2$ (benigna), significando que, para esse exemplo, ela não é importante para discriminar esses dois tipos de imagens, maligna e benigna, e portanto, o algoritmo Omega a remove do vetor de características.

7.2.3 Mineração de Regras de Associação

O método IDEA emprega o algoritmo Apriori para minerar as regras de associação. A saída do algoritmo Omega e as palavras-chave dos laudos do conjunto de imagens de treino são

submetidas ao algoritmo Apriori. Uma restrição, que limita a ocorrência das palavras-chave à cabeça das regras é adicionada ao processo de mineração. Os valores de confiança mínima utilizados devem ser altos, maiores que 97%. O valor do limite mínimo de confiança em 97% foi encontrado empiricamente, onde regras com confiança menores que essa produziram resultados não confiáveis no método IDEA.

O corpo das regras mineradas é composto pelos índices das características e seus intervalos. Um exemplo de regra minerada nesse passo do método IDEA é:

$3[0.1-0.3] \rightarrow$ *massa maligna* $(s = 0.05, c = 1.0)$

Essa regra indica que as imagens tendo a 3^a característica no intervalo $[0.1-0.3]$ tendem a ser imagens de massa maligna. Os valores 0.05 e 1.0 indicam, respectivamente, o suporte e a confiança da regra. O suporte de 0.05 significa que 5% das imagens de treino têm o valor da 3^a característica no intervalo $[0.1-0.3]$ e também são massas malignas. O valor de confiança 1.0 indica que 100% das imagens que têm o valor da 3^a característica no intervalo $[0.1-0.3]$ são massas malignas. As regras mineradas são utilizadas como entrada do algoritmo ACE, que é o único algoritmo utilizado na fase de teste do método IDEA.

7.2.4 O Algoritmo ACE

Antes de discutir o algoritmo ACE (**A**ssociative **C**lassifier **E**ngine), é necessário a definição de alguns termos.

Definição 7.1 *Uma imagem* **satisfaz** *uma regra, se os valores de suas características atendem todo o corpo da regra;*

Definição 7.2 *Uma imagem* **satisfaz parcialmente** *uma regra, se suas características atendem parte do corpo dessa regra;*

Definição 7.3 *Uma imagem* **não satisfaz** *uma regra, se suas características não atendem nenhuma parte do corpo da regra.*

ACE é um novo algoritmo apto a retornar múltiplas palavras-chave ao processar uma imagem de teste. O algoritmo ACE armazena todos os conjuntos de palavras-chave, aqui chamados de *itemsets*, que pertencem à cabeça das regras de treino em uma estrutura de dados. Seja, $M(h)$ o número de regras que o *itemset h satisfaz*; $P(h)$ o número de regras que o *itemset h satisfaz parcialmente*; e $N(h)$ o número de regras que o *itemset h não satisfaz*. O algoritmo ACE, usa uma medida chamada *convicção* (*conv*), para determinar se um *itemset* irá ou não ser adicionado ao diagnóstico sugerido de uma imagem. A *convicção* (*conv*) de um *itemset* é dada pela equação:

$$conv(h) = \frac{3M(h) + P(h)}{3M(h) + P(h) + N(h)} \tag{7.1}$$

A *convicção* indica o nível de certeza que um *itemset h* pertence ao diagnóstico final da imagem dado por um especialista. Um *itemset h* é retornado pelo algoritmo ACE no diagnóstico sugerido *se* a seguinte condição for satisfeita:

Condição 7.4 $M(h) \geq 1$ ***e*** $conv(h) \geq conv_{min}$

Quanto maior o valor da *convicção*, maior a confiança de que *h* pertence ao diagnóstico final da imagem. Um limiar de mínima *convicção* $conv_{min}(0 \leq conv_{min} \leq 1)$ é empregado para limitar a ocorrência de um *itemset* no diagnóstico sugerido pelo método. Se $conv_{min} = 0$, todos os *itemsets* que *satisfazem* pelo menos uma regra são retornados. A figura 7.3 mostra um exemplo de funcionamento do algoritmo ACE.

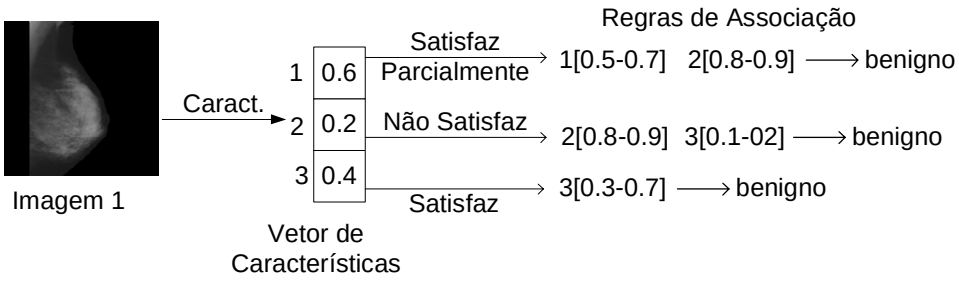


Figura 7.3: Exemplo de funcionamento do algoritmo ACE.

No exemplo da figura 7.3, $M(h) = 1$, $P(h) = 1$ e $N(h) = 1$ para o *itemset* $h = \{benigno\}$. Assim, se $\frac{4}{5} \geq conv_{min}$, o *itemset* $h = \{benigno\}$ é retornado pelo algoritmo como integrante do laudo da imagem de teste (Imagem 1), caso contrário ele é descartado.

ACE é detalhado no algoritmo 5. Como é mostrado nos estudos de caso, ACE é bastante adequado para gerar sugestões de diagnóstico em imagens médicas.

Esta seção apresentou o método IDEA. Embora a aplicação do método descrita nesta tese é análise de imagens médicas, o problema foi descrito genericamente, de maneira que se torne fácil a aplicação do mesmo para outros campos de pesquisas relacionados.

7.3 Experimentos

Aqui são apresentados dois estudos de caso realizados para validar o método IDEA na tarefa de sugerir diagnósticos para imagens médicas. Os experimentos foram feitos empregando 10% das imagens para teste e o restante das imagens para treino. Os parâmetros do algoritmo Omega

Algoritmo 5: Algoritmo ACE.

Dados: Vetor de características F da imagem de teste, conjunto de regras de associação S, e $conv_{min}$
Resultado: Conjunto de palavras-chave K

para *cada regra* $s \in S$ *na forma corpo* $\rightarrow$ *cabeça* **faça**
  para *cada itemset* $h \in$ *cabeça* **faça**
    se *corpo* ***satisfaz*** F **então**
      $M(h)++$;
    fim
    senão se *corpo* ***satisfaz parcialmente*** F **então**
      $P(h)++$;
    fim
    senão
      $N(h)++$;
    fim
  fim
fim
para *cada regra* $s \in S$ *da forma corpo* $\rightarrow$ *cabeça* **faça**
  para *cada itemset* $h \in$ *cabeça* **faça**
    se $M(h) \geq 1$ ***e*** $conv(h) = \frac{3M(h)+P(h)}{3M(h)+P(h)+N(h)} \geq conv_{min}$ **então**
      Adicione h in K;
    fim
  fim
fim
retorna K

foram definidos como $H_{min} = 2$, $\zeta_{max} = 0.2$ e $\zeta_{G_{max}} = 0.3$, que são parâmetros ótimos do algoritmo. Os valores de suporte mínimo *minsup=0.005* e confiança mínima *minconf=1.0* foram utilizados como entrada para o algoritmo Apriori. O valor de $conv_{min} = 0$, que maximiza o número de palavras-chave retornados pelo algoritmo ACE, foi empregado como parâmetro de entrada do mesmo. Os parâmetros dos algoritmos permaneceram os mesmos para os dois estudos de caso aqui apresentados.

7.3.1 Estudo de Caso 1

Neste estudo de caso é utilizada a base *ROI446*, que é composta de 446 imagens de regiões de interesse (ROIs) compreendendo tecidos tumorais de mamografias. Os ROIs dessa base foram obtidos do sistema "*the Breast Imaging Reporting and Data System*" do Departamento de Radiologioa da Universidade de Viena (www.birads.at). Esses ROIs estão disponíveis na Internet e são usados para treinar novos estudantes de radiologia na tarefa de diagnóstico de mamografias. Cada ROI possui um diagnóstico associado composto das seguintes partes:

- **Morfologia**: massa (circunscrita, indistinta, especular); distorção arquitetural; densidade

assimétrica, calcificação (amorfa, pleomorfa, linear, benigna);

- **BI-RADS**: seis níveis (0-5);

- **Histologia**: cisto, fibrose, tecido gorduroso, etc. (total de 25 palavras-chave).

A escala BI-RADS (*Breast Imaging Reporting and Data System*), resumida na tabela 7.1, foi desenvolvida pelo Colégio Americano de Radiologia para padronizar os laudos e os procedimentos de diagnóstico de mamografias.

Tabela 7.1: Níveis de BI-RADS.

value	description
0	Necessidade de outra avaliação da imagem.
1	Negativo.
2	Achado benigno.
3	Provavelmente achado benigno. Acompanhamento em pequenos intervalos de tempo.
4	Suspeita de anormalidade. Recomendação de biopsia.
5	Alta indicação de malignidade. Ação adequada deve ser tomada.

No passo de extração de características, as imagens foram segmentadas e características de textura, forma e cor foram extraídas das regiões segmentadas. O processo de segmentação foi feito eliminando inicialmente as regiões com nível de cinza menor que 0.14 (em uma escala de cinza [0-1]) e aplicando a técnica de segmentação de Otsu [Otsu, 1979] na imagem resultante. A figura 7.4 mostra um exemplo de imagem segmentada. As características apresentadas na tabela 7.2 foram extraídas das regiões segmentadas para compor o vetor de 15 características que representa as imagens.

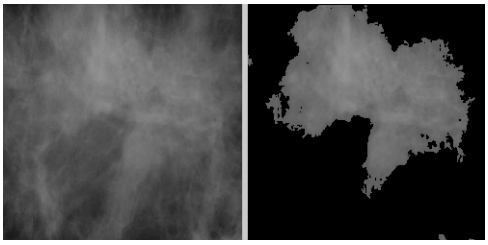


Figura 7.4: Exemplo de segmentação: imagem original (esquerda) e imagem segmentada (direita).

O algoritmo Omega foi aplicado sobre os vetores de características das imagens de treino, discretizando os valores das características e removendo a 13^a característica. Isso significa que a 13^a característica é a menos discriminante do vetor de características. A saída do algoritmo

Tabela 7.2: Características e suas posições no vetor de características.

feature	**position**
média de intensidade, contraste	1-2
rugosidade, distorção	3-4
uniformidade, entropia	5-6
momentos invariantes	7 to 13
media do histograma, desvio padrão	14-15

Omega foi submetida ao algoritmo Apriori onde 662 regras foram mineradas. Um exemplo de regra minerada nesse passo do método IDEA é:

$1[167.03 - 169.1] \rightarrow$ *carcinoma ductal invasivo (IDC)* $(s = 0.01, c = 1.0)$

Essa regra significa que as imagens tendo o valor da 1^a característica (média de intensidade) no intervalo fechado [167.03-169.1] tendem a ser imagens de *carcinoma ductal invasivo* (IDC), com suporte de 0.01 (1% das imagens de treino são IDC tendo o valor da 1^a característica no intervalo [167.03-169.1]) e confiança 1.0 (todas as images tendo o valor da 1^a característica no intervalo [167.03-169.1] são imagens de IDC).

As regras de associação geradas e as imagens de teste foram submetidas ao algoritmo ACE que sugeriu diagnósticos (conjunto de palavras-chave) para cada imagem de teste. O diagnóstico sugerido pelo algoritmo ACE foi comparado com o diagnóstico real das imagens dado por especialistas e por resultados de biopsias.

Para validar o método IDEA na tarefa de determinar o nível de BI-RADS ele foi comparado com outros três classificadores bem conhecidos:o **C4.5** [Quinlan, 1993]; o **Naive Bayes** [John & Langley, 1995]; e o 1- vizinho mais próximo (**1NN**) [Aha & Kibler, 1991]. A tabela 7.3 mostra os resultados. Como a escala BI-RADS tem uma separação nebulosa entre dois níveis consecutivos, mesmo para um experiente radiologista, tendo um alto grau de interseção e similaridade entre níveis consecutivos, foi considerado correto se o nível de BI-RADS sugerido pelos métodos fosse o mesmo ou adjacente ao nível anotado pelo radiologista no laudo.

Tabela 7.3: Comparação entre o método IDEA e os classificadores C4.5, *Naive Bayes* e 1NN na tarefa determinar o nível de BI-RADS das imagens de teste da base *ROI446*.

Medida	**IDEA**	**C4.5**	***Naive Bayes***	**1NN**
Precisão	**96.7%**	95.5%	86.7%	71.1
Sensibilidade	**91.3%**	85.7%	71.4%	50.0
Especificidade	**71.4%**	100.0%	77.8%	33.3

Note que o método IDEA produz os valores mais altos de precisão e sensibilidade. Uma vantagem do método IDEA é que, mantendo o mesmo custo computacional, ele permite sugerir um conjunto de palavras-chave no diagnóstico de uma imagem de teste, incluindo informações de morfologia e histologia no diagnóstico. Os outros classificadores também podem retornar um conjunto de palavras-chave. Entretanto, para cada palavra-chave retornada um novo modelo

de aprendizado deve ser construído, aumentando a complexidade e o custo computacional. Por essa razão, os resultados alcançados pelo método IDEA são mais factíveis de serem alcançados utilizando regras de associação do que usando outras técnicas de mineração.

A precisão obtida pelo método IDEA ao determinar a **morfologia** das ROIs foi **91.3%**. Os valores de precisão obtidos pelo método IDEA (96.7% para BI-RADS e 91.3% para morfologia) indicam que as características empregadas representam melhor o nível de BI-RADS do que as propriedades morfológicas da imagem.

Um protótipo, chamado sistema IDEA, incorporando o método IDEA foi implementado. O sistema IDEA também foi experimentado por dois radiologistas, que demonstraram interesse em utilizar o mesmo em seu dia-a-dia. Eles reportaram que o sistema indicou lesões que eles não viram em uma análise inicial. A figura 7.5 mostra uma tela do sistema IDEA ao analisar uma imagem. O sistema mostra a *convicção* de cada palavra-chave do diagnóstico entre parênteses. A *convicção* indica o nível de certeza que a respectiva palavra-chave irá pertencer ao diagnóstico final da imagem dado por um radiologista.

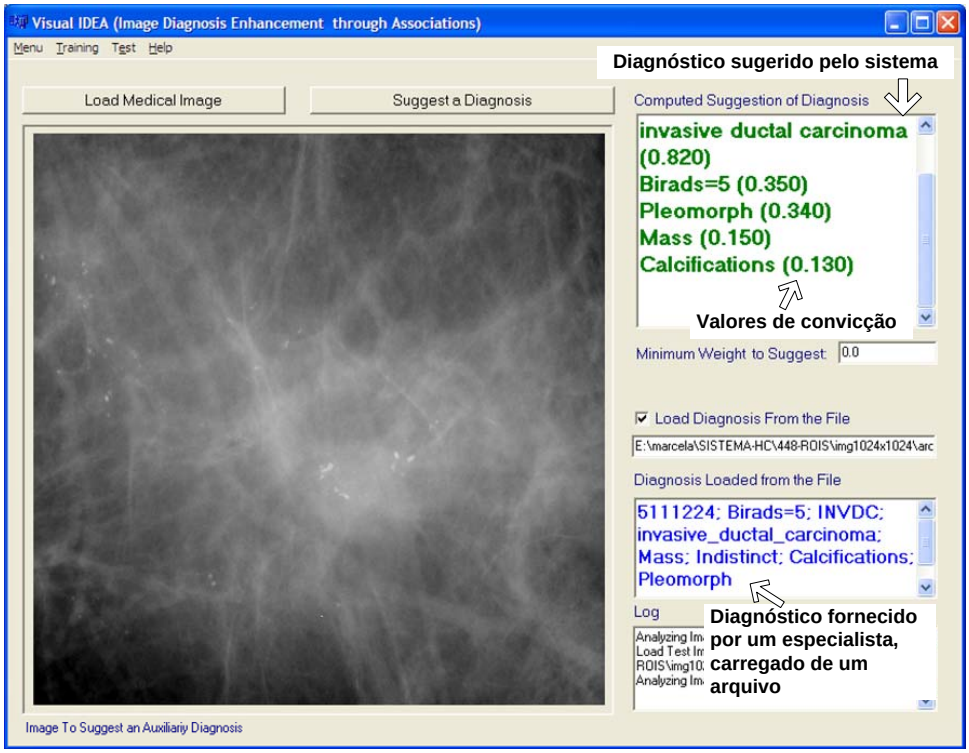


Figura 7.5: Tela do sistema IDEA.

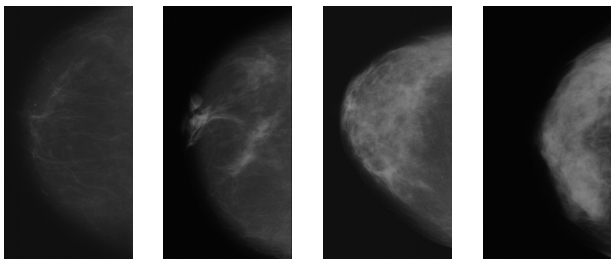
7.3.2 Estudo de Caso 2

Neste estudo de caso é utilizada a base *Mamografia1080*. A base *Mamografia1080* é composta de 1080 mamografias colhidas no Hospital das Clínicas da Universidade de São Paulo em Ribeirão Preto. As mamografias são classificadas em 4 níveis de densidade de acordo com a tabela

7.4. A figura 7.6 ilustra exemplos de imagens desta base. A densidade da mama é um fator de risco muito importante para o desenvolvimento do câncer de mama e também é um dos fatores que mais influencia na confiança de um diagnóstico feito por imagem.

Tabela 7.4: Os quatros níveis de densidade de mamografias e sua distribuição na base Mamografia1080.

nível	significado	número de imagens
1	gordurosa	362
2	parcialmente gordurosa	446
3	parcialmente densa	200
4	densa	72



nível 1 nível 2 nível 3 nível 4

Figura 7.6: Exemplos de imagens da base *Mamografia1080*. As imagens da esquerda para direita correspondem respectivamente a mamogramas dos níveis 1, 2, 3, e 4 de densidade.

Neste estudo de caso, as imagens são representadas pelo vetor de características proposto por [Kinoshita et al., 2007], compondo um vetor de 65 características, incluindo forma, tamanho da mama e distribuição de tecido fibroglandular. Essas características são sumarizadas na tabela 7.5. Esse vetor de característica é apontado na literatura como sendo um dos mais adequados para discriminar mamografias de diferentes densidades.

Tabela 7.5: Características e suas posições no vetor de características usado para representar a base Mamografia1080.

característica	posição
Granularidade	1-10
Forma	11-13
Momentos de Hu	14-27
Características de Haralick	28-55
Histograma	56-65

A análise visual de mamografias por radiologistas é uma tarefa subjetiva e sofre de alto grau de variabilidade. Assim, aqui é considerado correto se a densidade apontada pelo método avaliado é igual ou adjacente à densidade anotada pelo radiologista. Os mesmos passos do estudo de caso anterior foram utilizados aqui.

O algoritmo Omega reduziu 14% do vetor de características, mantendo somente 56 das 65 características originais do vetor de características. As características removidas pelo algoritmo Omega estão apresentadas na tabela 7.6.

Tabela 7.6: Características removidas do vetor de características ilustrado na tabela 7.5

Características	Posições Removidas
Momentos de Hu	18,19,20
Características de Haralick	25,26,27,36,50
Histograma	64

Além das palavras-chave relativas ao nível de densidade de mamografias, os diagnósticos sugeridos pelo método IDEA também contemplam informações referentes ao lado da mama (esquerdo ou direito) e ao tipo de visão (médio-lateral oblíqua ou crânio-caudal).

O algoritmo Apriori minerou 30996 regras de associação sobre a saída do algoritmo Omega. Um exemplo de regra minerada nesse passo é:

$15[0.0059 - 0.0064] \rightarrow$ *densidade=2, médio-lateral oblíqua* ($s = 0.005, 1.0$)

Essa regra indica que as imagens tendo o valor da 15^a característica (o segundo momento de Hu) no intervalo fechado [0.0059-0.0064] tendem a ser imagens de mamografias médio-lateral oblíqua com densidade 2 (parcialmente gordurosa), com suporte de 0.005 (0.5% das imagens de treino são imagens de mamografias medio-lateral oblíqua parcialmente gordurosas com o valor da 15^a característica no intervalo [0.0059-0.0064]) e confiança de 1.0 (todas as mamografias tendo o valor da 15^a característica no intervalo [0.0059-0.0064] são médio-lateral oblíquas parcialmente gordurosas).

Na fase de teste, as regras de associação e as características das imagens de teste foram submetidas ao algoritmo ACE.

Para validar o método IDEA, ele foi comparado com os classificadores C4.5, *Naive Bayes* e 1NN na tarefa de determinar o nível de densidade das mamografias. Os resultados desse experimento são apresentados na tabela 7.7.

Tabela 7.7: Comparação entre o método IDEA e os classificadores C4.5, *Naive Bayes* e 1NN na tarefa de determinar o nível de densidade das mamografias da base Mamografia1080.

	IDEA	*C4.5*	*Naive Bayes*	*1NN*
Precisão	**94.8**	92.6	78.7	90.7

Note que o método IDEA leva aos maiores valores de precisão (ver tabela 7.7). Além disso, o método IDEA sugere um conjunto de palavras-chave. Também foi medida a habilidade do método IDEA em distinguir: (a) lado da mama (mama direita ou mama esquerda); (b) tipo de visão (médio-lateral oblíqua ou crânio-caudal); e (c) o nível de densidade. Os valores de precisão obtidos nessas três partes do diagnóstico foram:

Lado: ***73.5%****; Visão:* ***85.1%****; Densidade:* ***94.8%****.*

A figura 7.7 mostra uma tela do sistema IDEA com os resultados do mesmo ao analisar a imagem apresentada na esquerda da tela. Observe que o sistema sugere as densidades 1 e 2 para a imagem analisada. Entretanto, a densidade 2 tem um valor mais alto de *convicção*, indicando que é mais provável que a imagem seja da densidade 2, mas o radiologista deve considerar esses dois valores de densidade. Note que a densidade anotada pelo radiologista no laudo é 2, ou seja, aquela que o sistema aponta um maior valor de *convicção*. Esse exemplo ilustra o funcionamento do sistema IDEA que pode sugerir mais de uma hipótese diagnóstica para o radiologista considerar, sendo importante o radiologista considerar todas as hipóteses sugeridas. No entanto, a hipótese com maior convicção deve ser analisada primeiro.

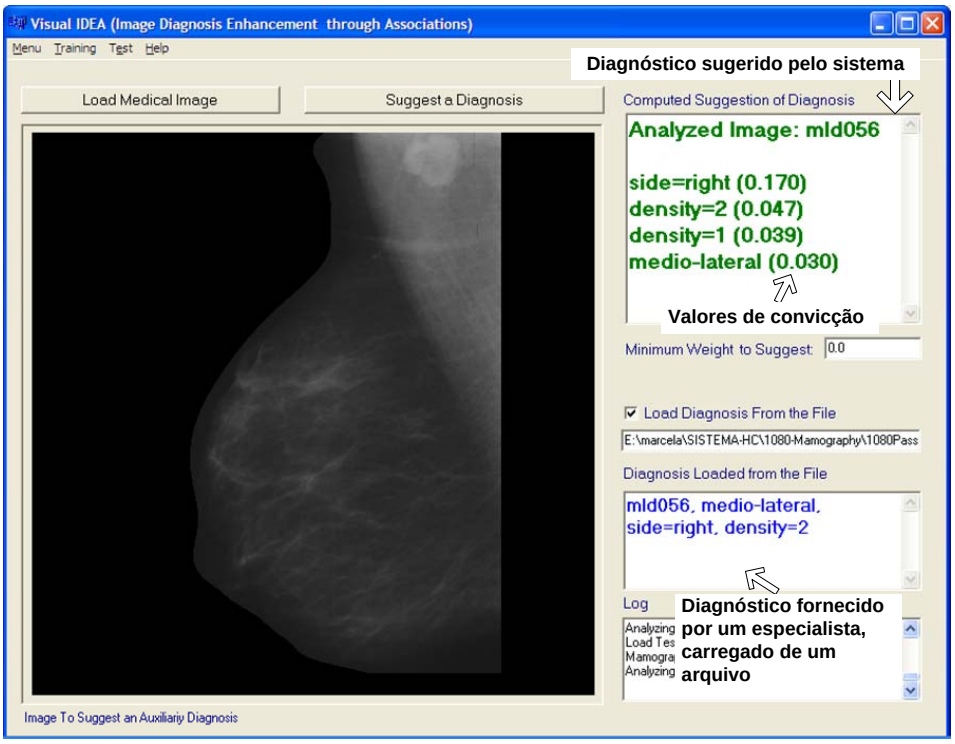


Figura 7.7: Tela do sistema IDEA.

Os resultados apresentados nesta seção são bastante promissores, pois mostram que a mineração de regras de associação pode ser utilizadas com sucesso para apoiar sistemas CAD, produzindo taxas de erro menores do que as técnicas tradicionais utilizadas.

7.4 Um Breve Histórico

Durante a elaboração desta tese, antes do método IDEA, foi desenvolvido o método SuGAR para o auxílio ao diagnóstico de imagens médicas com base na mineração de regras de associação. O método SuGAR [Ribeiro et al., 2007] foi uma tentativa inicial de desenvolvimento

de um método para auxílio ao diagnóstico com base na mineração de regras de associação. O método SuGAR é detalhado no apêndice C desta tese.

O método IDEA surgiu como uma evolução do método SuGAR. As principais diferenças em relação ao método SuGAR são descritas a seguir. No método IDEA, o extrator de características empregado varia de acordo com o tipo de imagens e o tipo de análise requerido, enquanto no SuGAR, um extrator de características único, composto por descritores de textura de Haralick, é utilizado para extrair características para todos os tipos de imagens e análises. No IDEA, o algoritmo Omega é utilizado, enquanto no SuGAR, o algoritmo PreSAGe é utilizado na fase de pré-processamento. No IDEA, a medida de interesse *convicção* é utilizada na fase de teste para a sugestão de diagnóstico, enquanto no SuGAR nenhuma medida de interesse é utilizada.

7.5 Considerações Finais

Neste capítulo foi apresentado um método para auxílio ao diagnóstico de imagens médicas baseado na mineração de regras de associação. Esse método emprega dois algoritmos novos Omega e ACE. O algoritmo Omega executa duas tarefas em um único passo: discretização e seleção de características. O algoritmo ACE gera sugestões de diagnóstico, associando um conjunto de palavras-chave a uma imagem de teste. Os resultados dos experimentos realizados com bases reais mostram que o método IDEA atinge valores de precisão e sensibilidade altos quando comparados com os algoritmos C4.5, *Naive Bayes* e *1NN*. Em adição, o método IDEA sugere um conjunto de palavras-chave com o mesmo custo computacional, enquanto os outros métodos que empregam classificação necessitam construir um modelo de aprendizado para cada palavra-chave retornada no diagnóstico, aumentando muito o esforço computacional. Por esta razão, os resultados obtidos pelo método IDEA são mais factíveis de serem obtidos empregando regras de associação do que empregando outras técnicas. Além disso, radiologistas que fizeram um uso inicial do sistema demostraram a aceitação do mesmo, expondo interesse de empregar o sistema para auxiliá-los no seu trabalho cotidiano.

Parte III

Conclusões e Trabalhos Futuros

Capítulo 8

Conclusões e Trabalhos Futuros

8.1 Considerações Iniciais

Atualmente, a mineração de imagens tem sido foco de muitas pesquisas no campo de mineração de dados e recuperação de informações [Stockman & Shapiro, 2001]. Um dos principais desafios desse campo é como efetivamente relacionar características de baixo nível, automaticamente extraídas das imagens através dos algoritmos de processamento de imagens, com características de alto nível, relativas à semântica e a interpretação humana dessas imagens. A mineração de regras de associação foi aplicada com sucesso em outras áreas, como negócios, e pode ser usada para revelar padrões interessantes relacionando características de baixo e alto nível das imagens.

Em medicina dois tipos de sistemas médicos estão se tornando amplamente utilizados: os sistemas CBIR *Content-based Image Retrieval* e os sistemas CAD *Computer Aided Diagnosis*. O propósito dos sistemas CAD é melhorar a consistência da interpretação das imagens com o uso do computador como uma segunda opinião para o radiologista. Semelhante aos sistemas CAD, os sistemas CBIR usam a informação extraída da própria imagem para representá-la, entretanto, seu principal propósito é recuperar imagens semelhantes. Analisando imagens similares, seus laudos e casos similares, o radiologista pode aumentar a precisão e a velocidade de análise das imagens. Além disso, os sistemas CAD e CBIR têm se mostrado bastante importantes no ensino e aprendizado de medicina.

Nesta tese, a mineração de regras de associação foi empregada para suportar esses dois tipos de sistemas médicos: CBIR e CAD. Comparações entre o uso de regras de associação e outras técnicas de mineração foram apresentadas nos experimentos realizados. Os resultados desses experimentos mostram que a mineração de regras de associação, em grande parte dos casos, leva a melhores resultados do que as outras técnicas de mineração testadas. Esses resultados vão ao encontro da afirmação feita por Holte [Holte, 1993] de que as técnicas mais simples

de mineração funcionam bem na maioria dos casos, se os parâmetros são convenientemente configurados.

8.2 Principais Contribuições

Nesta tese foram propostos métodos baseados em regras de associação para melhorar a precisão da busca por conteúdo e auxílio ao diagnóstico de imagens médicas.

Para melhorar a precisão da busca por conteúdo em imagens médicas foi proposto o algoritmo StARMiner, que minera regras de associação estatísticas relacionando as características mais representativas das imagens com suas classes. Essas regras de associação são utilizadas para reduzir o "gap semântico" entre as características e sua interpretação. As regras mineradas pelo algoritmo StARMiner são utilizadas para selecionar e ponderar os vetores de características utilizados para representar as imagens, reduzindo a dimensionalidade e aumentando a precisão das buscas por conteúdo em imagens médicas.

Para auxílio ao diagnóstico, foi proposto o método IDEA baseado na mineração de regras de associação em imagens médicas. O método IDEA sugere um conjunto de palavras-chave para compor o diagnóstico para uma dada imagem de teste. O diagnóstico sugerido pelo sistema IDEA pode ser utilizado para acelerar e trazer maior confiança ao trabalho de diagnosticar imagens por radiologistas. Para suportar o método IDEA, dois novos algoritmos foram propostos: Omega e ACE. O algoritmo Omega realiza simultaneamente discretização e seleção de características sendo bastante adequado para pré-processar as características das imagens médicas para a tarefa de associação. O algoritmo ACE é um classificador associativo que retorna múltiplas palavras-chave para compor o diagnóstico de uma imagem de teste. ACE usa medida de interesse chamada *convicção* que indica o quão provável é a ocorrência de uma determinada palavra-chave no laudo final da imagem dado pelo radiologista.

As técnicas desenvolvidas nesta tese foram comparadas com métodos tradicionais da literatura atingindo, na maioria das vezes, um resultado melhor que as demais técnicas testadas.

O desenvolvimento desta tese permitiu também que a aluna trabalhasse em assuntos correlatos ao foco desta tese:

- desenvolvimento e validação de técnicas de segmentação e extração de características;
- aumento da precisão de consultas utilizando realimentação de relevância;
- determinação de funções de distância adequadas para execução de consultas por similaridade em determinados tipos de imagens médicas;
- análise de fluxos de dados e mineração de regras de associação para a análise de dados agrometeorológicos.

No entanto, considera-se que a maior contribuição desta tese foi a prova de que regras de associação podem ser empregadas com sucesso para suportar sistemas de auxílio ao diagnóstico e busca por conteúdo em imagens.

8.3 Publicações

Considera-se também que as publicações alcançadas foram contribuições muito importantes desta tese. Essas publicações são listadas a seguir.

Periódicos Internacionais

[1] Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., C.; Marques, P. M. A. "An Association Rule-based Method to Support Medical Image Diagnosis with Efficiency". Special Issue on Multimedia Data Mining of IEEE Transactions on Multimedia, vol. 10 (2), p. 277-285, 2008.

[2] Ribeiro, M. X.; Bugatti, P. H.; Traina, Agma J. M.; Traina Jr, C.; . Marques, P. M. A.; Rosa, N. A. "Supporting Content-Based Image Retrieval and Computer-Aided Diagnosis systems with Association Rule-Based Techniques", Special issue on Knowledge Discovery in Medicine of Elsevier's Data & Knowledge Engineering Journal (DKE), 20 pags (artigo aceito, a ser publicado em 2009).

[3] Traina, A. J. M.; Traina Jr, C.; Ciferri, C. D. A. ; Ribeiro, M. X.; Marques, P. M. A. "How to Cope with the Performance Gap in Content-based Image Retrieval Systems". International Journal of Healthcare Information Systems and Informatics, 4(1), 47-67, January-March 2009.

[4] Balan, A. G. R.; Traina, A.J.M.; Ribeiro, M.X.; Azevedo, P. M. A.; Traina Jr., C. "Smart Histogram Analysis Applied to the Skull-stripping Problem in T1-weighted MRI", artigo submetido (na 2^a revisão) ao periódico IEEE Transactions on Medical Imaging (TMI), 10 pags.

Capítulo de Livro Internacional

[1] Ribeiro, M. X.; Balan, André G. R.; Felipe, J. C; Traina, A. J. M; Traina Jr., C., Mining Complex Data, "Mining Statistical Association Rules to Select the Most Relevant Medical Image Features", Mining Complex Data, Studies in Computational Intelligence, Springer, vol. 165/2009, 113-131, 2009.

Conferências Internacionais - Artigos Completos

[1] Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., C.; Rosa, N. A.; Marques, P. M. A. "How To Improve Medical Image Diagnosis through Association Rules: The IDEA Method". The 21th IEEE International Symposium on Computer-Based Medical Systems (CBMS), Jyväskylä, Finland, 2008, p. 266-271 (*Runner-up of the Best student paper*).

[2] Bugatti, P. H.; Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., C. "Content-based Retrieval

of Medical Images by Continuous Feature Selection". The 21th IEEE International Symposium on Computer-Based Medical Systems (CBMS), Jyväskylä, Finland, 2008, p. 272-277.

[3] Ribeiro, M. X. ; Ferreira, M. R. P ; Traina, A. J. M. ; Traina, C. . "Data Pre-processing: A new algorithm for Feature Selection and Data Discretization". The International Conference on Soft Computing as Transdisciplinary Science and Technology (ACM/IEEE CSTST'2008), 2008, Paris, France, 2008, p. 252-257.

[4] Ribeiro, M. X.; Traina, A. J. M.; Balan, A. G. R.; Traina Jr., C.; Marques, P. M. A. "SuGAR: A Framework to Support Mammogram Diagnosis".The 20th IEEE International Symposium on Computer-Based Medical Systems, Maribor, Slovenia, June 2007, p. 47-52.

[5] Balan, Andre G. R.; Traina, A. J. M.; Ribeiro, Marcela Xavier; Marques, Paulo M. d. A.; Traina Jr., C. "HEAD: the Human Encephalon Automatic Delimiter". The 20th IEEE International Symposium on Computer-Based Medical Systems, Maribor, Slovenia, June 2007, p. 171-176.

[6] Ribeiro, M. X.; Marques, J.; Traina, A. J. M.; Traina Jr., C. "Statistical Association Rules and Relevance Feedback: Powerful Allies to Improve the Retrieval of Medical Images". The 19th IEEE International Symposium on Computer-Based Medical Systems, Salt Lake City, Utah, USA, June 2006, p. 887 - 892.

[7] Felipe, J. C.; Traina, A. J. M.; Traina Jr., C.; Sousa, E. P. M.; Ribeiro, M. X. "Effective Shape-based Retrieval and Classification of Mammograms". The 21st Annual ACM Symposium on Applied Computing (SAC 2006), Dijon, France, 2006, p. 250-255.

Conferência Internacional - Resumo Estendido

[1] Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., C. "A new Algorithm for Data Discretization and Feature Selection". 23rd Annual ACM Symposium on Applied Computing, Fortaleza (CE), Brazil, 2008, p. 953-954.

Workshops Internacionais - Artigos Completos

[1] Silva, C. W.; Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., C. "Employing Wavelet Transforms to Support Content-Based Retrieval of Medical Images". In: 8th International Workshop on Pattern Recognition in Information Systems (PRIS 2008), 2008, Barcelona, p. 19-28.

[2] Ribeiro, M. X.; Balan, Andre G. R.; Traina, A. J. M.; Traina Jr., C. "Enhancing Medical Image Retrieval through Association Rules". In the Workshop of Content-based Image Retrieval for Biomedical Image Archives: Achievements, Problems, and Prospects of the 10th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2007, p. 11-20.

[3] Sousa, E. P. M.; Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., C. "Tracking the Intrinsic Dimension of Evolving Data Streams to Update Association Rules". 3rd International Workshop

on Knowledge Discovery from Data Streams junto ao the 23th International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, June, 2006, p. 1-10.

[4] Ribeiro, M. X.; Balan, Andre G. R.; Felipe, J. C.; Traina, A. J. M.; Traina Jr., C. "Mining Statistical Association Rules to Select the Most Relevant Medical Image Features". First International Workshop on Mining Complex Data (IEEE MCD'05) junto ao IEEE International Conference on Data Mining (ICDM'2005), Houston, USA, November 2005, p.91-98.

[5] Felipe, J. C.; Olioti, Jonatas B.; Traina, A. J. M.; Ribeiro, M. X.; Sousa, Elaine P. M. de; Traina Jr., C. "A Low-cost Approach for Effective Shape-based Retrieval and Classification of Medical Images". The First IEEE International Workshop on Multimedia Information Processing and Retrieval (IEEE-MIPR 2005) junto ao International Symposium on Multimedia (IEEE ISM'2005), Irvine, California, USA, 2005, p. 565-570.

Conferências Nacionais - Artigos Completos

[1] Silva, C. Y. V. W; Bugatti, P. H.; Ribeiro, M. X.; Traina, A. J. M.; Traina Jr., "Improving CBIR Using Feature Extraction Based on Wavelet Transform", XIV Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia2008), Vila Velha, ES, 2008, p.1-8.

[2] Ribeiro, M. X.; Marques, J.; Traina, A. J. M.; Traina Jr., C. "Combatendo os Pesadelos da Busca Por Conteúdo em Imagens Médicas". X Congresso Brasileiro de Informática em Saúde, Florianópolis -SC, outubro 2006, p. 1454-1459.

Conferência Nacional - Resumo Estendido

[1] Ribeiro, M. X.; Traina, A. J. M. "Usando Regras de Associação Estatísticas para a Seleção de Características Relevantes em Imagens Médicas". II Simpósio de Instrumentação e Imagens Médicas, São Pedro - SP, Outubro 2005, p. 60-61.

Workshops Nacionais - Artigos Completos

[1] Romani, L. A. S.; Traina, A. J. M.; Ribeiro, M. X.; Sousa, E. P. M.; Zullo Jr., J.; Traina Jr., C. "Aplicação de Técnicas de Mineração em Dados Climáticos e de Satélite para Auxiliar no Acompanhamento de Safras Agrícolas". IV Workshop em Algoritmos e Aplicações de Mineração de Dados em conjunto com XXIII Simpósio Brasileiro de Banco de Dados (WAMD'2008), Campinas - SP, Outubro de 2008, p. 1-6.

[2] Ribeiro, M. X.; Silva, C. W.; Felipe, J. C.; Balan, A. G. R.; Traina, A. J. M.; Traina Jr., C. "Apoiando a Busca por Conteúdo em Imagens Médicas através da Mineração de Regras de Associação Estatística". II Workshop em Algoritmos e Aplicacões de Mineração de Dados do XX Simpósio Brasileiro de Banco de Dados (WAAMD'2006), Florianópolis -SC, outubro 2006, p. 27-24.

[3] Ribeiro, M. X.; Felipe, J. C.; Traina, A. J. M. "Seleção de Atributos Relevantes para

Busca por Similaridade e Classificação de Imagens Médicas". I Workshop de Visão Computacional, Piracicaba, SP, p.44-47.

Workshop Nacional - Resumo Estendido

[1] Ribeiro, M. X.; Marques, J.; Traina, A. J. M.; Traina Jr., C. "Maximizando a Precisão de Buscas por Conteúdo em Imagens Médicas através da Mineração de Regras de Associação e Realimentação de Relevância". VI Workshop de Informática Médica, Vila Velha, ES, junho 2006, p.1-2.

8.4 Trabalhos Futuros

Os principais direcionamentos de trabalho futuro para este trabalho de pesquisa são listados a seguir:

- Avaliar o ganho de precisão alcançado pelos radiologistas com o uso do sistema IDEA através de construção de curvas ROC;
- Avaliar o ganho de tempo alcançado pelos radiologistas com o uso do sistema IDEA;
- Adicionar o método IDEA no sistema cbPACS e disponibilizá-lo para uso no hospital e melhorar sua usabilidade;
- Aplicar os métodos propostos a outros tipos e bases de imagens médicas;
- Utilizar outros extratores, discretizadores, seletores de características e classificadores para efetuar comparações com o método desenvolvido;
- Aplicar/desenvolver métodos de mineração de texto ao método IDEA para trabalhar com laudos completos e não somente com conjunto de palavras-chave;
- Desenvolver métodos de mineração de padrões seqüenciais, considerando seqüências de imagens e laudos e sua evolução a partir do tempo, visando antecipar o surgimento de alguma anomalia.

Referências Bibliográficas

[Abraham et al., 2006] Abraham, R., B.Simha, J., e Iyengar, S. (2006). A comparative analysis of discretization methods for medical datamining with naive bayesian classifier. In *9th International Conference on Information Technology*, pp. 235–236.

[Aggarwal, 2001] Aggarwal, C. C. (2001). On the effects of dimensionality reduction on high dimensional similarity search. In *Symposium on Principles of Database Systems (ACM PODS)*, pp. 1–11, Santa Barbara, CA.

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., e Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD International Conference on Management of Data*, v. 1, pp. 207–216.

[Agrawal & Srikant, 1994] Agrawal, R. e Srikant, R. (1994). Fast algorithms for mining association rules. In *International Conference on Very Large Databases (VLDB)*, pp. 487–499, Santiago de Chile, Chile.

[Aha & Kibler, 1991] Aha, D. e Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, (6):37–66.

[Al Aghbari & Al Haj, 2006] Al Aghbari, Z. e Al Haj, R. (2006). Hill-manipulation: An effective algorithm for color image segmentation. *Image and Vision Computing*, 24(8):894–903.

[Antonie et al., 2003] Antonie, M.-L., Zaiane, O. R., e Coman, A. (2003). Associative classifiers for medical images. In *LNAI 2797, MMCD*, pp. 68–83. Springer-Verlag.

[Armato III et al., 2001] Armato III, S. G., Giger, M. L., e MacMahon, H. (2001). Automated detection of lung nodules in ct scans: Preliminary results. *Medical Physics*, 28(8):1552–1561.

[Asuncion & Newman, 2007] Asuncion, A. e Newman, D. (2007). Uci machine learning repository (http://www.ics.uci.edu/ mlearn/mlrepository.html). University of California, Department of Information and Computer Science, Irvine, CA.

[Aumann & Lindell, 1999] Aumann, Y. e Lindell, Y. (1999). A statistical theory for quantitative association rules. In *The fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 261–270, San Diego, California, United States.

[Baeg & Kehtarnavaz, 2000] Baeg, S. e Kehtarnavaz, N. (2000). Texture based classification of mass abnormalities in mammograms. In *13th IEEE Symposium on Computer-Based Medical Systems (CBMS'00)*, pp. 163–168, Houston, TX.

[Baeza-Yates & Ribeiro-Neto, 1999] Baeza-Yates, R. A. e Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval.* Addison-Wesley, Wokingham, UK.

[Balan, 2007] Balan, A. G. R. (2007). *Métodos adaptativos de segmentação aplicados a recuperação de imagens por conteúdo*. Tese de doutorado, Universidade de São Paulo.

[Balan et al., 2007] Balan, A. G. R., Traina, A. J. M., Ribeiro, M. X., Marques, P. M. A., e Traina-Jr., C. (2007). Head: The human encephalon automatic delimiter. In *20th IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, pp. 171–176, Maribor, Slovenia.

[Balan et al., 2005] Balan, A. G. R., Traina, A. J. M., Traina Jr., C., e Marques, P. M. d. A. (2005). Fractal analysis of image textures for indexing and retrieval by content. In *18th IEEE Intl. Symposium on Computer-Based Medical Systems - CBMS*, pp. 581–586, Dublin, Ireland.

[Barioni, 2006] Barioni, M. C. N. (2006). *Operações de consulta por similaridade em grandes bases de dados complexos*. PhD thesis, Universidade de São Paulo.

[Beyer et al., 1999] Beyer, K., Godstein, J., Ramakrishnan, R., e Shaft, U. (1999). When is "nearest neighbor"meaningful? In *International Conference on Database Theory (ICDT)*, v. 1540, pp. 217–235, Jerusalem, Israel.

[Brin et al., 1997] Brin, S., Motwani, R., Ullman, J. D., e Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD International Conference on Management of Data*, pp. 255–264, Tucson, Arizona, USA.

[Bueno, 2002] Bueno, J. M. (2002). *Suporte à Recuperação de Imagens Médicas baseada em Conteúdo através de Histogramas Métricos*. Tese de doutorado, Universidade de São Paulo.

[Bugatti, 2008] Bugatti, P. H. (2008). *Análise da Influência de Funções de Distância para o Processamento Perceptual de Consultas por Similaridade em Sistemas PACS*. Dissertação de mestrado, Universidade de São Paulo.

[Bugatti et al., 2008] Bugatti, P. H., Ribeiro, M. X., Traina, A. J. M., e Jr, C. T. (2008). Content-based retrieval of medical images by continuous feature selection. In *The 21th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2008)*, pp. 272–277, Jyväskylä, Finland.

[Buhmann et al., 2007] Buhmann, S., Herzog, P., Liang, J., Wolf, M., Salganicoff, M., Kirchhoff, C., Reiser, M., e Becker, C. H. (2007). Clinical evaluation of a computer-aided diagnosis (cad) prototype for the detection of pulmonary embolism. *Academic Radiology*, 14(6):651–658.

[Cardie, 1993] Cardie, C. (1993). Using decision trees to improve case-based learning. In *10th International Conference on Machine Learning*, pp. 25–32.

[Cerquides & Mantaras, 1997] Cerquides, J. e Mantaras, L. d. (1997). Proposal and empirical comparison of a parallelizable distance-based discretization method. In *3rd International Conference on Knowledge Discovery and Data Mining*, pp. 139–142, California.

[Cheng et al., 2006] Cheng, T.-H., Wei, C.-P., e Tseng, V. S. (2006). Feature selection for medical data mining: Comparisons of expert judgment. In *the 19th IEEE International Symposium on Computer-based Medical Systems*, pp. 165–170, Salt Lake City, Utah, USA.

[Ciaccia et al., 1998] Ciaccia, P., Patella, M., e Zezula, P. (1998). A cost model for similarity queries in metric spaces. In *ACM Symposium on Principles of Database Systems (PODS)*, pp. 59–68, Seattle, Washington.

[Comer & Delp, 2000] Comer, M. L. e Delp, E. J. (2000). The em/mpm algorithm for segmentation of textured images: Analysis and further experimental results. *IEEE Transactions on Image Processing*, 9(10):1731–1744.

[Comer et al., 1996] Comer, M. L., Liu, S., e Delp, E. J. (1996). Statistical segmentation of mammograms. In *3rd International Workshop on Digital Mammography*, pp. 475–478.

[Costa & Jr., 2001] Costa, L. F. d. e Jr., R. M. C. (2001). *Shape Analysis and Classification - Theory and Practice*. CRC Press, Boca Raton, CA.

[Deserno et al., 2008] Deserno, T. M., Antani, S., e Long, R. (2008). Ontology of gaps in content-based image retrieval. *Journal of Digital Imaging*, 1(1):1–14.

[Doi, 2005] Doi, K. (2005). Current status and future potential of computer-aided diagnosis in medical imaging. *The British Journal of Radiology*, 78:S3–S19.

[Doi, 2007] Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review,current status and future potential. *Computerized Medical Imaging and Graphics*, 31(1):198–211.

[Dougherty et al., 1995] Dougherty, J., Kohavi, R., e Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pp. 194–202.

[Dy et al., 2003] Dy, J., Brodley, C., Kak, A., Broderick, L.S. A4 Broderick, L., e Aisen, A.M. A5 Aisen, A. (2003). Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):373–378.

[Egecioglu et al., 2004] Egecioglu, O., Ferhatosmanoglu, H., e Ogras, U. (2004). Dimensionality reduction and similarity computation by inner-product approximations. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(6):714–726.

[Faloutsos & Kamel, 1994] Faloutsos, C. e Kamel, I. (1994). Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *ACM Symposium on Principles of Database Systems (PODS)*, pp. 4–13, Minneapolis, MN.

[Fayyad & Irani, 1993] Fayyad, U. M. e Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Artificial Intelligence*, pp. 1022–1029.

[Fayyad et al., 1996] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., e Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA.

[Felipe et al., 2006] Felipe, J. C., Ribeiro, M. X., Sousa, E. P. M. d., Traina, A. J. M., e Traina Jr., C. (2006). Effective shape-based retrieval and classification of mammograms. In *21st Annual ACM Symposium on Applied Computing (SAC 2006)*, pp. 250–255, Dijon, France.

[Felipe et al., 2003] Felipe, J. C., Traina, A. J. M., e Traina-Jr., C. (2003). Retrieval by content of medical images using texture for tissue identification. In *16th IEEE Symposium on Computer-Based Medical Systems*, pp. 175–180, New York.

[Ferreira et al., 2008] Ferreira, C., Torres, R., Gonçalves, M., e Fan, W. (2008). Image retrieval with relevance feedback based on genetic programming. In *XXIII Simpósio Brasileiro de Banco de Dados - SBBD*, pp. 1–15, Campinas, SP, Brazil.

[Flickner & alli, 1995] Flickner, M. e alli, e. (1995). Query by image and video content: The qbic system. *IEEE Computer*, 28(9):23–32. September 1995.

[Fowlkes et al., 2008] Fowlkes, C. C., Hendriks, C. L. L., Keränen, S. V., Weber, G. H., Rübel, O., Huang, M.-Y., Chatoor, S., DePace, A. H., Simirenko, L., Henriquez, C., Beaton, A., Weiszmann, R., Celniker, S., Hamann, B., Knowles, D. W., Biggin, M. D., Eisen, M. B., e Malik, J. (2008). A quantitative spatiotemporal atlas of gene expression in the drosophila blastoderm. *Cell*, 133(2):364–374.

[Frawley et al., 1991] Frawley, W. J., Piatetsky-Shapiro, G., e Matheus, C. J. (1991). Knowledge discovery in databases: An overview. pp. 1–27. AAAI/MIT Press.

[Freer & Ulissey, 2001] Freer, T. W. e Ulissey, M. J. (2001). Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. *Radiology*, 220(3):781–786.

[Gerhardinger, 2006] Gerhardinger, L. C. (2006). *Segmentação de imagens e validação de classes por abordagem estocástica*. Dissertação de mestrado, Universidade de São Paulo.

[Giger, 2000] Giger, M. L. (2000). Computer-aided diagnosis in medical imaging – a new era in image interpretation. *World Medical Journal*, 1(1):75–79.

[Gonzalez & Woods, 2008] Gonzalez, R. C. e Woods, R. E. (2008). *Digital Image Processing*. Prentice Hall, 5 edition.

[Guttman, 1984] Guttman, A. (1984). R-tree : A dynamic index structure for spatial searching. In *ACM SIGMOD International Conference on Management of Data*, pp. 47–57, Boston, MA.

[Hadjiiski et al., 2004] Hadjiiski, L., Chan, H.-P., Sahiner, B., Helvie, M. A., Roubidoux, M. A., Blane, C., Paramagul, C., Petrick, N., Bailey, J., Klein, K., Foster, M., Patterson, S., Adler, D., Nees, A., e Shen, J. (2004). Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: An roc study. *Radiology*, 233(1):255–265.

[Han & Kamber, 2000] Han, J. e Kamber, M. (2000). *Data Mining - Concepts and Techniques*. Morgan Kaufmann Publishers, New York, 1st edition edition.

[Han et al., 2000] Han, J., Pei, J., e Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pp. 1–12, Dallas, Texas, USA.

[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., e Distein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3:610–621.

[Heath et al., 2001] Heath, M., Bowyer, K., Kopans, D., Moore, R., e Kegelmeyer, W. P. (2001). The digital database for screening mammography. In *the Fifth International Workshop on Digital Mammography*, pp. 212–218. Medical Physics Publishing.

[Hendriks et al., 2006] Hendriks, C. L. L., Keränen, S. V., Fowlkes, C. C., Simirenko, L., Weber, G. H., DePace, A. H., Henriquez, C., Kaszuba, D. W., Hamann, B., Eisen, M. B., Malik, J., Sudar, D., Biggin, M. D., e Knowles, D. W. (2006). Three-dimensional morphology and gene expression in the drosophila blastoderm at cellular resolution i: data acquisition pipeline. *Genome Biology*, 7(12):R123.1–R123.20.

[Hipp et al., 2000] Hipp, J., Güntzer, U., e Nakhaeizadeh, G. (2000). Algorithms for association rule mining - a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64.

[Ho & Scott, 1997] Ho, K. M. e Scott, P. D. (1997). Zeta: A global method for discretization of continuous variables. In *Third International Conference on Knowledge Discovery and Data Mining*, pp. 191–194.

[Holte, 1993] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91.

[Houtsma & Swami, 1993] Houtsma, M. e Swami, A. N. (1993). Set-oriented mining of association rules. Research Report RJ 9567, IBM Almaden Research Center.

[Hsu et al., 2002] Hsu, W., Lee, M. L., e Zhang, J. (2002). Image mining: Trends and developments. *Journal of Intelligent Information Systems*, 19(1):7 – 23.

[Huang & Dai, 2003] Huang, P.-W. e Dai, S. K. (2003). Image retrieval by texture similarity. *Pattern Recognition Letters*, 36(3):665–679.

[Jackle et al., 1992] Jackle, H., Hoch, M., Pankratz, M., Gerwin, N., Sauer, F., e Bronner, G. (1992). Transcriptional control by drosophila gap genes. *Journal of Cell Science, Supplement*, 16:39–51.

[Jain, 1993] Jain, A. K. (1993). *Fundamentals of Digital Image Processing*. Prentice Hall.

[Jeong et al., 2007] Jeong, S., Kim, S.-W., e Choi, B.-U. (2007). Dimensionality reduction in high-dimensional space for multimedia information retrieval. In *18th International Conference on Database and Expert Systems Applications*, pp. 404–413, Regensburg, Germany.

[Jiang et al., 2001] Jiang, Y., Nishikawa, R. M., Schmidt, R. A., Toledano, A. Y., e Doi, K. (2001). Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology*, 220(3):787–794.

[John & Langley, 1995] John, G. H. e Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. pp. 338–345, San Mateo. Morgan Kaufmann.

[Kass et al., 1987] Kass, M., Witkin, A., e Terzopoulos, D. (1987). Snakes: Active countour models. *International Journal of Computer Vision*, 1(4):321–331.

[Katayama & Satoh, 1997] Katayama, N. e Satoh, S. (1997). The sr-tree: An index structure for high-dimensional nearest neighbor queries. In *ACM SIGMOD International Conference on Management of Data*, pp. 369–380, Tucson, Arizona, USA.

[Kazemzadeh & Sartipi, 2006] Kazemzadeh, R. S. e Sartipi, K. (2006). Incoporating data mining applications into clinical guildelines. In *19th IEEE Symposium on Computer-Based Medical Systems*, pp. 321–328, Salt Lake City, Utah, USA.

[Kerber, 1992] Kerber, R. (1992). Chimerge: Discretization of numeric attributes. In *10th International Conference on Artificial Intelligence*, pp. 123–128.

[Khotanzad & Hong, 1990] Khotanzad, A. e Hong, Y. H. (1990). Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(5):489–497.

[Kinoshita et al., 2007] Kinoshita, S. K., Azevedo-Marques, P. M. d., Jr, R. R. P., Rodrigues, J. A. H., e Rangayyan, R. M. (2007). Content-based retrieval of mammograms using visual features related to breast density patterns. *Journal of Digital Imaging*, 20(2):172–190.

[Kira & Rendell, 1992] Kira, K. e Rendell, L. A. (1992). A practical approach for feature selection. In *9th Intl. Conf. on Machine Learning*, pp. 249–256, Aberdeen, Scotland.

[Ko et al., 2000] Ko, B., Lee, H.-S., e Byun, H. (2000). Image retrieval using flexible image subblocks. In *ACM Symposium on Applied Computing (SAC)*, v. 2, pp. 574 – 578.

[Kobayashi & Doi, 1999] Kobayashi, T. e Doi, K. (1999). Effect of cad on radiologists' detection of lung nodules on chest radiographs. *Innervision*, 14:44–46.

[Kobayashi et al., 1996] Kobayashi, T., Xu, X.-W., MacMahon, H., CE, M., e K, D. (1996). Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology*, 199:843–848.

[Kononenko, 1994] Kononenko, I. (1994). Estimating attributes : Analysis and extension of relief. In *European Conference on Machine Learning* 171–182, p., Catania, Italy.

[Korn et al., 2001] Korn, F., Pagel, B.-U., e Faloutsos, C. (2001). On the 'dimensionality curse' and the 'self-similarity blessing'. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(1):96–111.

[Krishnapuram et al., 2004] Krishnapuram, R., Medasani, S., Jung, S.-H., Choi, Y.-S., e Balasubramaniam, R. (2004). Content-based image retrieval based on a fuzzy approach. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(10):1185–1199.

[Kurgan & Cios, 2004] Kurgan, L. A. e Cios, K. J. (2004). Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 16(2):145–153.

[Lawrence, 1993] Lawrence, P. A. (1993). *The Making of a Fly*, v. 1. Blackwell Scientific Publications, Cambridge, MA.

[Lin & Chen, 2008] Lin, H.-Y. e Chen, S.-Y. (2008). High indexing compression for spatial databases. In *IEEE 8th International Conference on Computer and Information Technology Workshops (CIT 2008)*, pp. 20–25.

[Lin et al., 1994] Lin, K.-I. D., Jagadish, H. V., e Faloutsos, C. (1994). The tv-tree: An index structure for high-dimensional data. *VLDB Journal*, 3(4):517–542.

[Liu et al., 2002] Liu, H., Hussain, F., Tan, C. L., e Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(1):393–423.

[Liu & Setiono, 1997] Liu, H. e Setiono, R. (1997). Feature selection via discretization. *Knowledge and Data Engineering*, 9(4):642–645.

[Liu & Wang, 2005] Liu, X. e Wang, H. (2005). A discretization algorithm based on a heterogeneity criterion. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(9):1166–1173.

[Liu et al., 2007] Liu, Y., Zhang, D., Lu, G., e Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition Letters*, 40:262–282.

[Ma et al., 1997] Ma, W., Deng, Y., e Manjunath, B. (1997). Tools for texture and color-based search of images. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging II*, v. 3016, pp. 496–507, San Jose, CA, USA.

[MacMahon et al., 1999] MacMahon, H., Engelmann, R., Behlen, F. M., Hoffmann, K. R., Ishida, T., Roe, C., Metz, C. E., e Doi, K. (1999). Computer-aided diagnosis of pulmonary nodules: Results of a large-scale observer test. *Radiology*, 213:723–726.

[Malcok et al., 2006] Malcok, M., Aslandogan, Y., e Yesildirek, A. (2006). Fractal dimension and similarity search in high-dimensional spatial databases. In *IEEE International Conference on Information Reuse and Integration (IRI 2006)*, pp. 380–384, Waikoloa, Hawaii, USA.

[Marques, 2001] Marques, P. M. d. A. (2001). Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, 34(5):285–293.

[Marques et al., 2008] Marques, P. M. d. A., Rosa, N. A., Traina, A. J. M., Jr, C. T., Kinoshita, S. K., e Rangayyan, R. M. (2008). Reducing the semantic gap in content-based image retrieval in mammography with relevance feedback and inclusion of expert knowledge. *International Journal of Computer Assisted Radiology and Surgery*, 3(1-2):123–130.

[McNicholas et al., 2008] McNicholas, P. D., Murphy, T. B., e O'Regan, M. (2008). Standardising the lift of an association rule. *Comput. Stat. Data Anal.*, 52(10):4712–4721.

[Molina et al., 2002] Molina, L. C., Belanche, L., e Nebot, n. (2002). Feature selection algorithms: A survey and experimental evaluation. In *IEEE International Conference on Data Mining (ICDM'02)*, pp. 306–404, Washington, DC, USA.

[Moreno & Furuie, 2006] Moreno, R. A. e Furuie, S. S. (2006). Biram: Sistema para recuperação de imagens por conteúdo. In *X Congresso Brasileiro de Informática em Saúde*, pp. 554–559.

[Morioka et al., 2005] Morioka, C. A., El-Saden, S., Pope, W., Duckwiler, G., Bui, A., e Kangarloo, H. (2005). Integration of his/ris clinical document with pacs image studies for neuroradiology. In *SPIE International Symposium on Medical Imaging*, San Diego, CA, USA.

[Mukherjea et al., 1999] Mukherjea, S., Hirata, K., e Hara, Y. (1999). AMORE: A World Wide Web Image Retrieval Engine. *World Wide Web*, 2(3):115–132.

[Muller et al., 2004] Muller, H., Michoux, N., Bandon, D., e Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23.

[Narendra & Fukunaga, 1977] Narendra, P. M. e Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions On Computer*, 26(9):917–922.

[Olukunle & Ehikioya, 2002] Olukunle, A. e Ehikioya, S. (2002). A fast algorithm for mining association rules in medical image data. In *IEEE CCECE*, pp. 1181–1187.

[Omiecinski, 2003] Omiecinski, E. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69.

[Ordonez et al., 2006] Ordonez, C., Ezquerra, N., e Santana, C. A. (2006). Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3):259–283.

[Ordonez & Omiecinski, 1999] Ordonez, C. e Omiecinski, E. (1999). Discovering association rules based on image content. *IEEE Forum ADL*, pp. 38–49.

[Otsu, 1979] Otsu, N. (1979). A thresholding selection method from graylevel histogram. *IEEE Transactions on Systems, Man and Cybernetics*, 9:62–66.

[Pagel et al., 2000] Pagel, B.-U., Korn, F., e Faloutsos, C. (2000). Deflating the dimensionality curse using multiple fractal dimensions. In *IEEE International Conference on Data Engineering (ICDE)*, pp. 589–598, San Diego, CA.

[Pan et al., 2005] Pan, H., Li, J., e Wei, Z. (2005). Mining interesting association rules in medical images. In *Advance Data Mining and Medical Applications (ADMA)*, pp. 598–609, Wuhan, China. Springer.

[Pass et al., 1996] Pass, G., Zabih, R., e Miller, J. (1996). Comparing images using color coherence vector. In *ACM Multimedia*, pp. 65–73, Boston, MA.

[Pentland et al., 1996] Pentland, A., Picard, R., e Sclaroff, S. (1996). Photobook: tools for content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254.

[Poonguzhali et al., 2007] Poonguzhali, S., Deepalakshmi, B., e Ravindran, G. (2007). Optimal feature selection and automatic classification of abnormal masses in ultrasound liver images. In *International Conference on Signal Processing, Communications and Networking*, pp. 503–506.

[Quek et al., 2003] Quek, S., Thng, C., Khoo, J., e Koh, W. (2003). Radiologists' detection of mammographic abnormalities with and without a computer-aided detection system. *Australasian Radiology*, 47(3):257–269.

[Quinlan, 1993] Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA.

[Rangayyan, 2005] Rangayyan, R. M. (2005). *Biomedical Image Analysis*, v. 1 of *Biomedical Engineering*. CRC Press, Florida, USA, 1 edition.

[Rangayyan et al., 2000] Rangayyan, R. M., Mudigonda, N., e Desautels, J. (2000). Boundary modelling and shape analysis methods for classification of mammographic masses. *Medical, Biological Engineering and Computing*, 38(1):487–496.

[Refaeilzadeh et al., 2007] Refaeilzadeh, P., Tang, L., e Liu, H. (2007). On comparison of feature selection algorithms. In *AAAI 2007 Workshop on Evaluation Methods for Machine Learning II*, Vancouver, Canada.

[Ribeiro et al., 2008a] Ribeiro, M. X. ., Ferreira, M. R. P., Traina Jr, C., e Traina, A. J. M. (2008a). Data pre-processing: A new algorithm for feature selection and data discretization. In *Fifth International Conference on Soft Computing as Transdisciplinary Science and Technology - ACM/IEEE (CSTST 2008)*, pp. 1–8, Cergy-Pontoise/Paris, France.

[Ribeiro et al., 2005a] Ribeiro, M. X., Balan, A. G. R., Felipe, J. C., Traina, A. J. M., e Traina Jr., C. (2005a). Mining statistical association rules to select the most relevant medical image features. In *First International Workshop on Mining Complex Data (IEEE MCD'05)*, pp. 91–98, Houston, USA.

[Ribeiro et al., 2005b] Ribeiro, M. X., Felipe, J. C., e Traina, A. J. M. (2005b). Seleção de atributos relevantes para busca por similaridade e classificação de imagens médicas. In *I Workshop de Visão Computacional*, pp. 44–47, Piracicaba, SP.

[Ribeiro et al., 2006a] Ribeiro, M. X., Marques, J., Traina, A. J. M., e Traina Jr., C. (2006a). Combatendo os pesadelos da busca por conteúdo em imagens médicas. In *X Congresso Brasileiro de Informática em Saúde*, pp. 1454–1459, Florianópolis, Santa Catarina.

[Ribeiro et al., 2006b] Ribeiro, M. X., Marques, J., Traina, A. J. M., e Traina Jr., C. (2006b). Maximizando a precisão de buscas por conteúdo em imagens médicas através da mineração de regras de associação e realimentação de relevância. In *VI Workshop de Informática Médica*, pp. 1 – 2 (CD–ROM), Vila Velha, ES.

[Ribeiro et al., 2006c] Ribeiro, M. X., Marques, J., Traina, A. J. M., e Traina-Jr, C. (2006c). Statistical association rules and relevance feedback: Powerful allies to improve the retrieval of medical images. In *19th IEEE International Symposium on Computer-Based Medical Systems*, pp. 887 – 892, Salt Lake City, Utah, USA.

[Ribeiro & Traina, 2005] Ribeiro, M. X. e Traina, A. J. M. (2005). Usando regras de associação estatísticas para a seleção de características relevantes em imagens médicas. In *II Simpósio de Instrumentação e Imagens Médicas*, pp. 60–61, São Pedro -SP.

[Ribeiro et al., 2007] Ribeiro, M. X., Traina, A. J. M., Balan, A. G. R., Jr., C. T., e Marques, P. M. A. (2007). Sugar: A framework to support mammogram diagnosis. In *20th IEEE International Symposium on Computer-Based Medical Systems*, pp. 47–52, Maribor, Slovenia.

[Ribeiro et al., 2008b] Ribeiro, M. X., Traina, A. J. M., Jr., C. T., Rosa, N. A., e Marques, P. M. A. (2008b). How to improve medical image diagnosis through association rules: The idea

method. In *The 21th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2008)*, pp. 1–6, Jyväskylä, Finland.

[Ribeiro et al., 2008c] Ribeiro, M. X., Traina, A. J. M., Traina, Jr., C., e Azevedo-Marques, P. M. A4 Azevedo-Marques, P. M. (2008c). An association rule-based method to support medical image diagnosis with efficiency. *IEEE Transactions on Multimedia*, 10(2):277–285.

[Ribeiro et al., 2008d] Ribeiro, M. X., Traina, A. J. M., e Traina Jr., C. (2008d). A new algorithm for data discretization and feature selection. In *23rd Annual ACM Symposium on Applied Computing*, pp. 1–2, Fortaleza, Ceará, Brazil.

[Ribeiro & Vieira, 2004] Ribeiro, M. X. e Vieira, M. T. P. (2004). A new approach for mining association rules in data warehouses. In *6th International Conference On Flexible Query Answering Systems*, v. 3055 of *Lecture Notes in Computer Science*, pp. 28–110, Lyon, France.

[Ribeiro et al., 2006d] Ribeiro, M. X., Watanabe, C. Y. V., Felipe, J. C., Balan, A. G. R., Traina, A. J. M., e Traina Jr., C. (2006d). Apoiando a busca por conteúdo em imagens médicas através da mineração de regras de associação estatística. In *II Workshop em Algoritmos e Aplicações de Mineração de Dados do XX Simpósio Brasileiro de Banco de Dados*, pp. 27–24, Florianópolis, Santa Catarina.

[Romani et al., 2008] Romani, L. A. S., Traina, A. J. M., Ribeiro, M. X., Souza, E. P. M., Zullo-Jr, J., e Jr., C. T. (2008). Aplicação de técnicas de mineração em dados climáticos e de satélite para auxiliar no acompanhamento de safras agrícolas. In *IV Workshop em Algoritmos e Aplicações de Mineração de Dados (WAAMD) em conjunto com XXIII o Simpósio Brasileiro de Banco de Dados (SBBD)*, pp. 1–6, Campinas, SP, Brazil.

[Rosa, 2002] Rosa, N. A. (2002). *Uma Abordagem Prática e Eficiente de Consultas por Similaridade para Suporte a Diagnóstico por Imagens*. Dissertação de mestrado, Universidade de São Paulo.

[Rosa, 2007] Rosa, N. A. (2007). *Inserção do conhecimento do especialista no processo de realimentação de relevância em recuperação de imagem por conteúdo: um estudo de viabilidade em mamografia*. Tese de doutorado, Universidade de São Paulo.

[Rosa et al., 2002] Rosa, N. A., Traina, A. J. M., e Traina Jr., C. (2002). Recuperação de imagens médicas por similaridade em um hospital universitário. In *2º Workshop de Informática Médica – WIM'2002 - in CD-ROM*, pp. 1–4, in CD–ROM, Gramado, RS.

[Rubner & Tomasi, 2001] Rubner, Y. e Tomasi, C. (2001). *Perceptual Metrics for Image Database Navigation*. The Kluwer Intl. Series in Engineering and Computer Science. Kluwer Academic Publishers.

[Savarese et al., 1995] Savarese, A., Omiecinski, E., e Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. In *Proc. of the 21st Conf. on Very Large Databases (VLDB'95)*, pp. 432–444.

[Siegel & Kolodner, 1999] Siegel, E. L. e Kolodner, R. M. (1999). *Filmless Radiology*. Springer Verlag, New York City, NY.

[Silva et al., 2008] Silva, C. Y. V. W., Bugatti, P. H., Ribeiro, M. X., Traina, A. J. M., e Traina Jr, C. (2008). Improving cbir using feature extraction based on wavelet transform. In *XIV Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia2008)*, pp. 1–8, Vila Velha, ES.

[Silva, 2007] Silva, C. Y. V. W. d. (2007). *Extração de características de imagens médicas utilizando wavelets para mineração de imagens e auxílio ao diagnóstico*. Dissertação de mestrado, USP.

[Silva, 2008] Silva, M. P. d. (2008). *Processamento de Consultas por Similaridade em Imagens Médicas Visando à Recuperação Perceptual Guiada pelo Usuário*. Qualificação de mestrado, Universidade de São Paulo.

[Sousa, 2006] Sousa, E. P. (2006). *Identificação de correlações usando a Teoria dos Fractais*. PhD thesis, Universidade de São Paulo.

[Sousa et al., 2006] Sousa, E. P. M. d., Ribeiro, M. X., Traina, A. J. M., e Traina Jr., C. (2006). Tracking the intrinsic dimension of evolving data streams to update association rules. In *3rd International Workshop on Knowledge Discovery from Data Streams in conjunction with the 23th International Conference on Machine Learning*, pp. 1–10 (CD–ROM), Pittsburgh, Pennsylvania, USA,.

[Sousa et al., 2002] Sousa, E. P. M. d., Traina Jr., C., Traina, A. J. M., e Faloutsos, C. (2002). How to use fractal dimension to find correlations between attributes. In *First Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches (in conjunction with 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)*, pp. 26–30, Edmonton, Alberta, Canada.

[Srikant & Agrawal, 1996] Srikant, R. e Agrawal, R. (1996). Mining quantitative association rules in large relational tables. In *ACM SIGMOD International Conference on Management of Data*, pp. 1–12, Montreal, Canada.

[Stockman & Shapiro, 2001] Stockman, G. e Shapiro, L. G. (2001). *Computer Vision*. Pearson Education.

[Thabtah, 2007] Thabtah, F. (2007). A review of associative classification mining. *Knowl. Eng. Rev.*, 22(1):37–65.

[Traina Jr. et al., 2007] Traina Jr., C., Santos Filho, R. F., Traina, A. J., Vieira, M. R., e Faloutsos, C. (2007). The omni-family of all-purpose access methods: a simple and effective way to make similarity search more efficient. *The VLDB Journal*, 16(4):483–505.

[Traina Jr. et al., 2005] Traina Jr., C., Sousa, E. P. M. d., e Traina, A. J. M. (2005). Using fractals in data mining. In *New Generation of Data Mining Applications*, v. 1 30p., p. Wiley/IEEE Press.

[Traina Jr. et al., 1997] Traina Jr., C., Traina, A. J. M., Santos, R. R. d., e Senzako, E. Y. (1997). A support system for content-based medical image retrieval in object oriented databases. *Journal of Medical Systems*, 21(6):339–352.

[Traina Jr. et al., 2002] Traina Jr., C., Traina, A. J. M., Santos Filho, R. F., e Faloutsos, C. (2002). How to improve the pruning ability of dynamic metric access methods. In *International Conference on Information and Knowledge Management (CIKM)*, pp. 219–226, McLean, VA, USA. ACM Press.

[Traina Jr. et al., 2000a] Traina Jr., C., Traina, A. J. M., Seeger, B., e Faloutsos, C. (2000a). Slim-trees: High performance metric trees minimizing overlap between nodes. In *International Conference on Extending Database Technology (EDBT)*, v. 1777 of *Lecture Notes in Computer Science*, pp. 51–65, Konstanz, Germany.

[Traina Jr. et al., 2000b] Traina Jr., C., Traina, A. J. M., Wu, L., e Faloutsos, C. (2000b). Fast feature selection using fractal dimension. In *Brazilian Symposium on Databases (SBBD)*, pp. 158–171, João Pessoa, PB.

[Veloso et al., 2003] Veloso, A., Jr., W. M., Parthasarathy, S., e Carvalho, M. d. (2003). Efficient, accurate and privacy-preserving data mining for frequent itemsets in distributed databases. In *XVIII Brazilian Symposium on Databases*, pp. 281–292, Manaus, AM.

[Vincent & Soile, 1991] Vincent, L. e Soile, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(6):583–598.

[Wang et al., 2004] Wang, X., Smith, M., e Rangayyan, R. (2004). Mammographic information analysis through association-rule mining. In *IEEE CCGEI 2004.*, pp. 1495–1498.

[Wang & Wu, 2005] Wang, Y. e Wu, X. (2005). Approximate inverse frequent itemset mining: Privacy, complexity, and approximation. In *the Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 482–489, Houston, TX.

[Wichert, 2008] Wichert, A. (2008). Subspace indexing for extremely high-dimensional cbir. In *International Workshop on Content-Based Multimedia Indexing (CBMI 2008)*, pp. 330–337.

[Yamamoto et al., 2008] Yamamoto, C. H., Oliveira, M. C. F., Rezende, S. O., e Nomelini, J. (2008). Including the user in the knowledge discovery loop: Interactive itemset-driven rule extraction. In *23rd. ACM Symposium on Applied Computing (SAC) - Multimedia and Visualization Track*, v. 2, pp. 1212–1217, Fortaleza, CE.

[Zaiane et al., 2002] Zaiane, O. R., Antonie, M.-L., e Coman, A. (2002). Mammography classification by an association rule-based classifier. In *The Third International Workshop on Multimedia Data Mining (MDM/KDD'2002)*, pp. 62–69, Edmonton, Alberta, Canada.

[Zaki et al., 1997] Zaki, M. J., Parthasarathy, S., Ogihara, M., e Li, W. (1997). New algorithms for fast discovery of association rules. In *the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 283–286, Newport Beach, CA, USA.

[Zhang et al., 2005] Zhang, C., Yang, Q., e Liu, B. (2005). Guest editors' introduction: Special section on intelligent data preparation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(9):1163–1165.

[Zhang & Su, 2002] Zhang, H.-J. e Su, Z. (2002). Relevance feedback in cbir. In *Sixth IFIP Working Conference on Visual Database Systems*, v. 216 of *IFIP Conference Proceedings*, pp. 21–35, Brisbane, Australia.

Parte IV

Apêndices

Apêndice A

O Projeto de Mineração de Embriões de *Drosophila*

Neste apêndice são descritas as etapas iniciais do projeto de mineração de padrões de expressão genética em imagens tridimensionais da espécie *Drosophila melanogaster*, popularmente conhecida como mosca da fruta. Este projeto teve início em setembro de 2008, em um estágio curto realizado pela candidata na Universidade *Carnegie Mellon* (CMU) em Pittsburgh - EUA, contando com a supervisão de dois renomeados pesquisadores desta instituição, prof. Dr. Christos Faloutsos e prof. Dr. Eric Xing.

O desenvolvimento da espécie *Drosophila melanogaster* pode ser descrito através de um sistema celular complexo e tridimensional que sofre alteração com o tempo, como conseqüência da multiplicação celular [Fowlkes et al., 2008]. Um dos desafios deste projeto é cruzar características visuais extraídas das imagens dos embriões com as informações sobre o tipo de gene que foi estimulado, o estado de desenvolvimento e outras informações semânticas, de maneira a mapear quais são as características visuais que descrevem uma determinada informação semântica. Além disso, outro desafio é simular o comportamento de um gene ao longo do tempo e as influências do mesmo sobre o comportamento dos demais genes.

A.1 Motivação

Um embrião animal pode ser considerado um conjunto de células tridimensionais que expressam a ação dos genes ao longo do tempo. É a ação dos genes que determina a diferenciação celular nos embriões [Lawrence, 1993].

A blastoderme da espécie *Drosophila melanogaster* tem sido utilizada como um modelo de estudo do comportamento genômico animal, por ser a espécie que melhor caracteriza os relacionamentos reguladores dos genes [Jackle et al., 1992]. Além disso, existem outras vantagens de utilizá-la como base de estudo para a compreensão e mapeamento dos genes [Lawrence, 1993], a saber:

- seu ciclo de vida é curto (dura de 10 a 14 dias);
- centenas de *drosophilas* podem ser mantidas em pequenos compartimentos;
- possui três cromossomos considerados "gigantes", que são utilizados como base de análise por apresentarem um comportamento padrão. O comportamento padrão desses cromossomos se deve principalmente por serem originários da célula da fêmea, não sendo afetados pelo material genético do macho.

Aqui, vale lembrar o conceito de cromossomos. Os cromossomos são longas seqüências de DNA (Ácido Desoxirribonucleico), que contém vários genes.

A.2 A Base de Dados

As imagens utilizadas neste projeto foram coletadas na Universidade de Berkeley, como parte do projeto *Berkeley Drosophila Transcription Network Project* - BDTNP (http://bdtnp.lbl.gov/FlyNet).

A base de dados utilizada no projeto contém imagens de 1282 embriões e descreve a expressão mRNA (ácido ribonucleico mensageiro) de 22 genes, em múltiplos períodos de tempo nos estágios de desenvolvimento do embrião 4d e 5. Apesar de existirem imagens de embriões com a expressão de mRNA de outros genes, esses outros genes, a princípio, não podem ser utilizadas como objeto de estudo, pois não possuem imagens na maioria dos períodos de desenvolvimento do embrião. O mRNA é produzido com base no modelo do DNA e é responsável por carregar a informação genética para os locais de síntese de proteínas nas células.

O embrião atinge o estágio de desenvolvimento 5 após duas horas e meia da fertilização. Nesse estágio o embrião de *drosophila* é formado por um conjunto de aproximadamente 6000

células envolvidas por uma membrana (blastoderme), que é a região mais importante de análise do embrião [Hendriks et al., 2006].

A base de dados é formada por imagens no formato ".lsm". As imagens desse tipo são obtidas por um tipo de microscópio de varredura a laser da marca Zeiss, que gera imagens tridimensionais do embrião. Cada imagem tridimensional é composta por aproximadamente uma centena de imagens bidimensionais do tipo ".tiff", colhidas em diferentes profundidades (fatias) do embrião. A figura A.1 ilustra um exemplo de imagem tridimensional de um embrião de *drosophila* no estágio 5. Na figura A.1 a imagem tridimensional do embrião é visualizada usando a ferramenta *Zeiss LSM Image Browser* (http://www.zeiss.de).

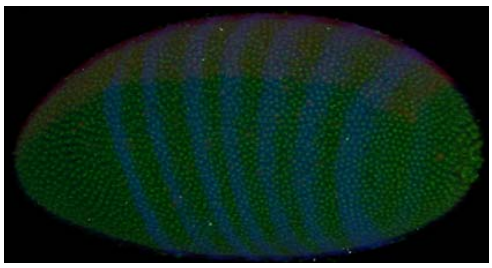


Figura A.1: Exemplo de imagem tridimensional de um embrião de *drosophila*

Uma imagem no formato ".lsm" ocupa aproximadamente 0.5 gigabyte de espaço de armazenamento, o que torna seu processamento bastante lento. Para simplificar o processamento da mesma, pesquisadores da Universidade de Berkeley desenvolveram um método para mapear as imagens tridimensionais em "*pointclouds*" (nuvem de pontos). Uma *pointcloud* é um arquivo texto que descreve o centro de massa das coordenadas dos núcleos das células na superfície do embrião e o nível de expressão mRNA de dois genes ao longo de cada núcleo [Hendriks et al., 2006]. Um dos genes descrito é chamado **gene de referência** (possui um comportamento padrão ao longo dos diferentes estágios) e o outro gene é chamado de **gene alvo** (gene cujo comportamento se deseja analisar). As *pointclouds* das imagens tridimensionais também foram disponibilizadas para serem utilizadas nesta pesquisa. A figura A.2 mostra a visualização da *pointcloud* correspondente ao embrião da figura A.1, onde a intensidade da cor está relacionada a atuação do gene "eve" (*even skipped*) na célula.

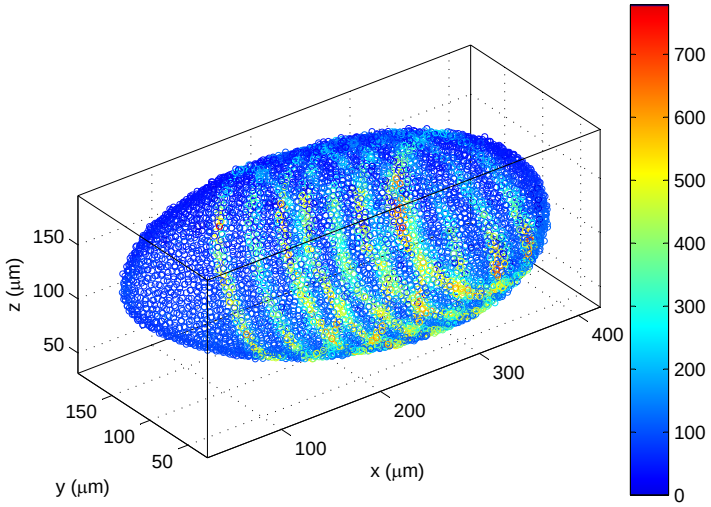


Figura A.2: Visualização da *pointcloud* correspondente ao embrião da imagem apresentada na figura A.1.

A.3 Obtendo as *Pointclouds*

Para extrair padrões das imagens tridimensionais dos embriões e de suas *pointclouds* primeiramente é necessário entender como ambas foram obtidas. As principais etapas do processo de aquisição das imagens e geração das *pointclouds*, de acordo com [Hendriks et al., 2006], são:

- Primeiramente os embriões são fixados e submetidos a uma tinta fluorescente para marcar os padrões de expressão do mRNA de dois genes (gene alvo e gene de referência). Um dos genes é o gene de referência que, em geral, é um gene bem comportado. É a análise do gene de referência que determina o estágio de desenvolvimento do embrião. Geralmente os genes de referência utilizados são os genes *eve*, *ftz* e *sna*;

- Uma vez marcada a expressão do mRNA dos dois genes e determinado o estágio de desenvolvimento, o microscópio é utilizado para produzir as imagens tridimensionais, não importando a orientação do embrião. Após esse procedimento, o embrião é descartado;

- As imagens tridimensionais são convertidas por processamento e análise de imagens em *pointclouds*;

- As *pointclouds* são analisadas usando várias técnicas de processamento de imagens para determinar a orientação do seus eixos anterior/posterior (a/p) e dorsal/ventral (d/v). Desse

modo, as *pointclouds* não estão todas na mesma orientação. Além disso, a localização e a quantidade da expressão dos genes é medida. Com as informações obtidas, metadados são gerados e adicionados aos arquivos das *pointclouds*;

- As imagens e as *pointclouds* são renderizadas e visualizadas utilizando um software específico.

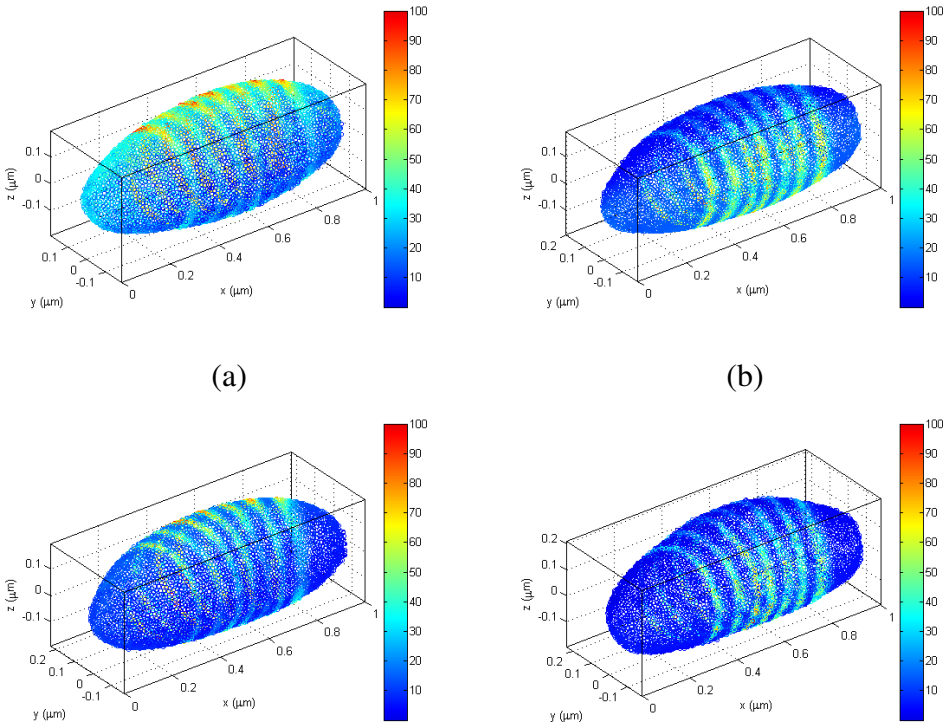
A.4 Alinhamento das imagens

Por requerer menor capacidade de processamento e armazenamento em disco, as *pointclouds*, ao invés das imagens tridimensionais, foram escolhidas para serem exploradas inicialmente nesta pesquisa.

O cabeçalho de uma *pointcloud* possui algumas informações importantes:

- os ângulos que as *pointclouds* devem ser rotacionadas de maneira a sempre alinhar seus eixos (a/p e d/v);
- o sub-estágio. Em geral um embrião possui por volta de uma centena de sub-estágios para um determinado estágio. A separação desses sub-estágios é nebulosa, sendo que os biólogos recomendam que as imagens sejam analisadas em grupos de aproximadamente 20 sub-estágios consecutivos;
- a qualidade (valores de 0 a 5). Os biólogos recomendam que somente *pointclouds* com qualidade igual ou superior a 4 sejam utilizadas para a extração de padrões;
- descreve qual canal de cor (*Cy3*, *Coumarin* e *Cytox Green*) está relacionado com a expressão de um determinado gene.

As *pointclouds* foram visualizadas e processadas utilizando o software Matlab. A figura A.3 mostra um exemplo da atividade do gene *eve* em embriões nos sub-estágios 25, 50, 75 e 100 do estágio 5 de desenvolvimento. As *pointclouds* visualizadas na figura A.3 são de qualidade 5 e foram previamente alinhadas. Observe que essas *pointclouds* possuem tamanhos, formatos e números de células diferentes.



(a) (b)

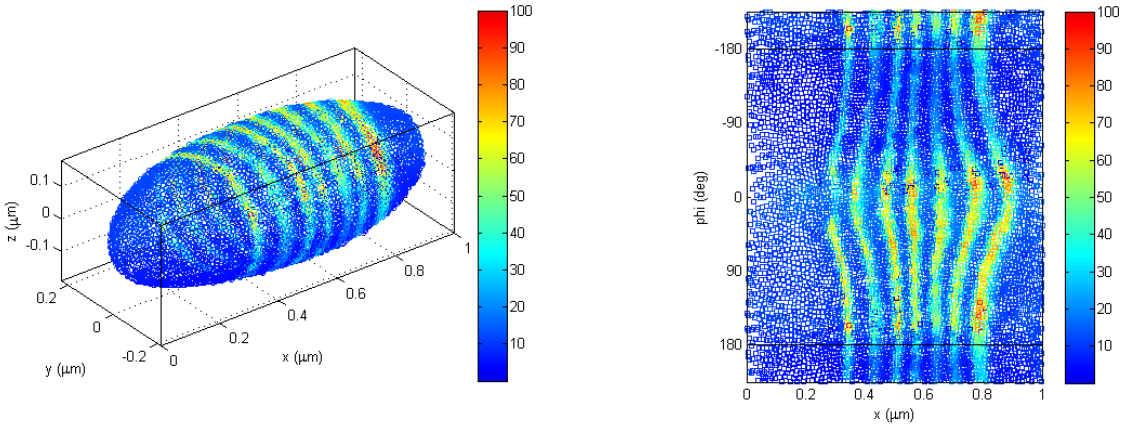
(c) (d)

Figura A.3: Visualização de *pointclouds* de qualidade 5 previamente alinhadas, ilustrando a atividade do gene *eve* nos sub-estágios 25 (a), 50 (b), 75 (c) e 100 (d) do estágio 5.

Uma das opções para facilitar o trabalho com as *pointclouds* é mapeá-las para um plano bidimensional. Isso pode ser feito através do mapeamento dos seus pontos para coordenadas cilíndricas. Nesse mapeamento, todos os pontos de um mesmo disco perpendicular ao eixo a/p de uma *pointcloud*, serão mapeados para a mesma reta vertical no plano. A principal limitação desse método é a ocorrência de muitos buracos nas extremidades do eixo a/p do embrião, representados pelo eixo x na figura A.4. A figura A.4 mostra uma *pointcloud* e sua projeção em coordenadas cilíndricas.

A.5 Próximos Passos

O primeiro objetivo deste projeto de pesquisa é descrever o comportamento dos genes ao longo do seu ciclo de vida, predizendo o comportamento de um gene em função de outro gene. Além disso, questões como “determine qual é o gene cujo comportamento mais se assemelha ao comportamento de um dado gene” e “dada uma imagem de um embrião com a expressão de



(a) (b)

Figura A.4: Exemplo de uma *pointcloud* (a) e sua projeção em coordenadas cilíndricas (b).

um gene, determinar em qual sub-estágio de vida o mesmo está". No entanto, para atingir os objetivos do projeto, algumas tarefas deverão ser realizadas:

- Registro das células dos embriões. Os embriões devem ser compatibilizados, de maneira a todos apresentarem o mesmo número de células, nas mesmas posições. Para isso deve ser determinado um processamento de mapeamento que não distorça as informações biológicas contidas nas imagens e/ou nas *pointclouds*;
- Determinação da representação da atividade de um gene nas células e seu movimento ao longo do tempo;
- Determinação da função de distância a ser usada para comparar seqüências de diferentes genes;
- Validação das técnicas desenvolvidas através da avaliação dos especialistas do domínio e do uso de classificadores.

Apêndice B

PreSAGe

Here, we describe an algorithm called PreSAGe (**Pre**-processing **S**olution for **A**ssociation rule **Ge**neration). PreSAGe is employed to perform the discretization of continuous values of the feature vector. The number of intervals found by it is used to select the relevant features. PreSage uses the concepts of "cut points" and "majority classes". ***Cut points*** are the limits of an interval of values. ***Majority class*** is the most frequent class of an interval.

B.1 The PreSAGe Algorithm

PreSAGe processes each feature separately. Let $f_{i,j}$ be the value of feature f_j in the feature vector i. PreSAGe uses a data structure that links each $f_{i,j}$ with the class c_i, where c_i is the class of image represented by the feature vector i. We refer to as instance I_i a pair ($f_{i,j}$, c_i). Let M_k be the majority class of an interval T_k and $|M_k|$, the number of occurrences of M_k in the interval T_k. Figure B.1 exemplifies the data structure used by PreSAGe algorithm.

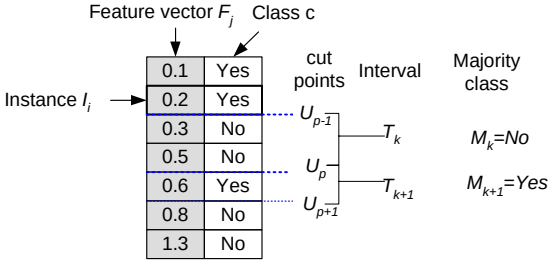


Figura B.1: Illustration of PreSAGe data structure.

When determining the data intervals, the algorithm PreSAGe creates a cut point u_p if:

Condition DA-1: The class label of the current instance I_i, $i > 1$, is different from the class label of the previous instance, i.e. $c_i \neq c_{i-1}$;

Condition DA-1 generates too many cut points, specially when working with noisy data. The larger the number of cut points is, the larger the number of intervals. Each interval represents an item in the process of mining association rules. The use of many items potentially generates a huge number of uninteresting rules, with low confidence. Hence, it is important to keep a small number of cut points and, consequently, a small number of items. PreSAGe removes unnecessary cut points in two stages. In the first stage it is considered the following condition:

Condition DA-2: The number of occurrences of the majority class in an interval T_k must be equal or greater than the *minperint* threshold, i.e. $|M_k| \geq minperint$.

In Condition DA-2, ***minperint*** is the restriction of the minimal allowed number of occurrences of the majority class in an interval. If Condition DA-2 is not satisfied by an interval, its right cut point is removed.

In the second stage, PreSAGe uses another condition, as follows:

Condition DA-3: The middle cut point of two consecutive intervals T_k and T_{k+1} is removed if $M_k = M_{k+1}$ **and** $\frac{|M_k|}{|T_k|} \geq mintofuse$ **and** $\frac{|M_{k+1}|}{|T_{k+1}|} \geq mintofuse$, where $|T_k|$ is the number of instances belonging to the interval T_k.

In Condition DA-3, ***mintofuse*** is used to control the minimum occurrence of the majority class in an interval. Condition DA-3 states that two consecutive intervals are fused if: they have the same majority class **and** if both have the fraction of occurrences of the majority class equal or greater than the *mintofuse* threshold. Figure B.2 presents an example of how PreSAGe works.

If an association rule mining algorithm (e.g Apriori [Agrawal & Srikant, 1994]) is applied to a dataset discretized by PreSAGe, rules of the form $F_j = T_k \rightarrow c$ can be mined. These rules indicate that images having the value of feature F_j in the interval T_k tend to be images of class c. This kind of rule is very useful for a classification task, where it is possible to infer the class of the images having the feature values.

Here we called the Discretization Approach (**DA**) approach, the feature selection promoted

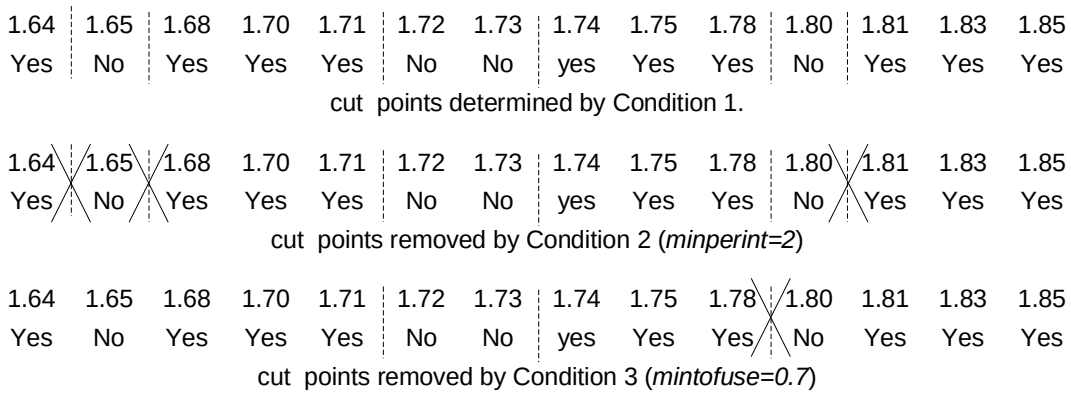


Figura B.2: Illustration of PreSAGe workflow.

using the PreSAGe algorithm. In the DA approach, a threshold ***nbrR*** is used to state the number of relevant features that should be returned by PreSAGe algorithm. The features are ranked according to the number of their intervals. Since the cut points are defined following the variation of the class label, the most relevant features are the ones presenting the smallest class variation, i.e., the ones generating fewer cut points. Therefore, the feature selection performed by DA approach is: *The features selected are the nbrR features that generates the smallest number of intervals by PreSAGe algorithm.*

DA is a very fast and straightforward approach that performs properly the task of feature selection. In our experiments, it presents the highest precision rates for similarity queries when compared to the original data and the other feature selection methods.

B.2 Experiments

Among several datasets we have considered, we show here results from two meaningful ones: the "Mammogram" and the "Heterogeneous" datasets, which we present as follows.

The Mammogram dataset is composed of 1,080 mammograms collected in the Clinical Hospital of University of Sao Paulo at Ribeirão Preto. The Mammogram dataset contains images classified in 4 levels of breast tissue density. In our experiments, images are represented by the feature set proposed in [Kinoshita et al., 2007], compounding a vector of 85 features. The visual analysis of mammograms by radiologists is a subjective task and suffers from a high degree of variability. Thus, it is a fair criterion to set as relevant not only the images of the same density

class of the query image, but also the images in the adjoining classes. This is the approach used here.

The Heterogeneous dataset consists of 704 medical images obtained from the same Clinical Hospital. It contains images obtained from Magnetic Resonance Imaging (MRI) divided in 8 classes.The feature vector considered for the Heterogeneous dataset is the same one used in [Balan et al., 2005], which is composed of 30 features.

Here, we call the Statistical Approach (**SA**) the approach that employs the StARMiner algorithm to mine statistical association rules from features of a training dataset. The mined rules are used to select the most relevant features. If a feature does not occur in a rule, it is removed from the feature vector, otherwise it is kept.

B.2.1 Case Study 1

The first experiment was performed using the Mammogram dataset divided in: training set (270 images) and test set (810 images).

StARMiner algorithm parameters were set to $\gamma_{min} = 0.9$, $\Delta\mu_{min} = 0.04$ and $\Delta\sigma_{max} = 0.8$. StARMiner algorithm was performed producing 48 rules and, consequently, selecting 48 features, what gives a reduction of 43% on the feature vector size.

In order to have a fair comparison among the algorithms, we set the number of features to be selected by PreSAGe (DA) as $nbrR = 48$, which is the same number of features returned by StARMiner (SA). PreSAGe algorithm was then run, having its parameters set to *maxperint*= 5 and *minfusion*= 0.8.

The feature vectors generated using the statistical approach (StARMiner) and the discretization approach (PreSAGe), were compared to the feature vector composed of the 48 most relevant features selected by Relief-F [Kira & Rendell, 1992]. The Precision and Recall (P&R) curve for each method, including the curve obtained by using the original 85-feature-vector, is presented in Figure B.3. Each curve represents an average P&R curve obtained by performing one similarity query for each image in the training set.

Figure B.3 shows that the results obtained using a more compact feature vector with 48 features are better than the results achieved using the original feature vector composed of 85

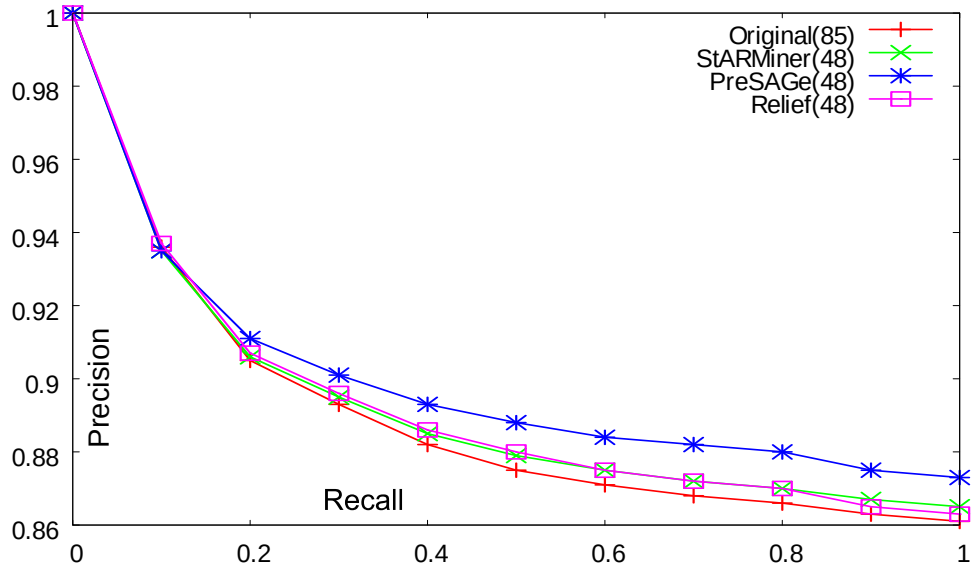


Figura B.3: P&R curves obtained for the Mammogram dataset.

features. The PreSAGe (the Discretization Approach) outperforms the other approaches, reaching the highest values of precision (always over 88%). Relief-F took 3.09 seconds to select the features, while Presage took 2.41 and StARMiner took 1.78 seconds (42% less time than Relief-F). In this experiment, the StARMiner algorithm performs similar to Relief-F approach, what is a good result, since Relief-F is a well established algorithm widely used to perform feature selection.

B.2.2 Case Study 2

The second experiment was performed using the "Heterogeneous" dataset divided in: training set (176 images) and test set (528 images).

StARMiner algorithm parameters were set to $\gamma_{min} = 0.98$, $\Delta\mu_{min} = 0.2$ and $\Delta\sigma_{max} = 0.13$, and the algorithm was executed over the training set, producing 21 rules and, consequently, selecting 21 features. Once again, we set the number of features to be returned as relevant by PreSAGe algorithm as $nbrR = 21$ (the same number of features returned by StARMiner). The parameters of PreSAGe algorithm were again set to *maxperint*= 5 and *minfusion*=0.8. The algorithms performed a reduction of 30% of the original feature vector size.

The feature vectors generated by StARMiner and PreSAGe algorithms, were again compa-

red to the vectors composed of the 21 most relevant features selected by Relief-F. The P&R curve obtained for each method, including the curve for the original feature vectors (30 features), is presented in Figure B.4.

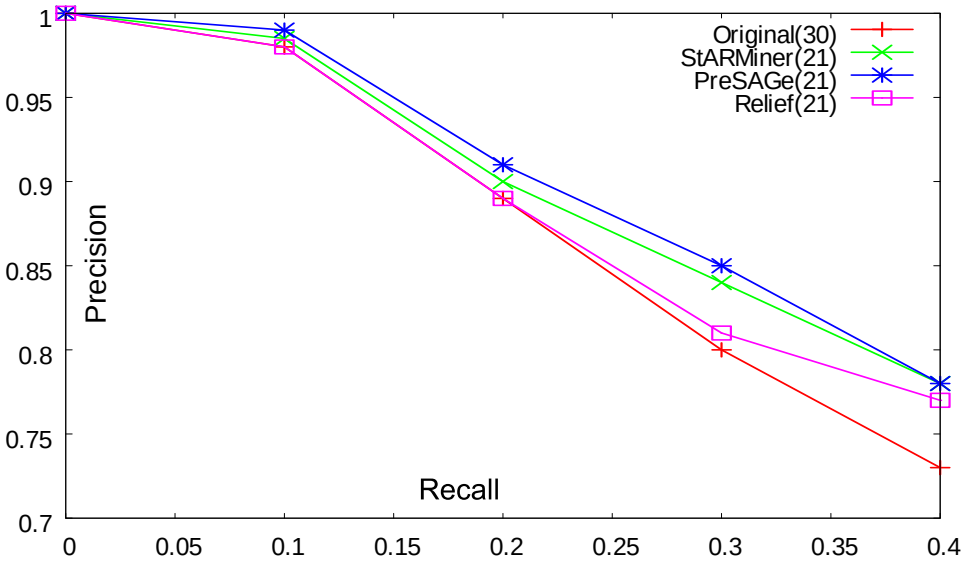
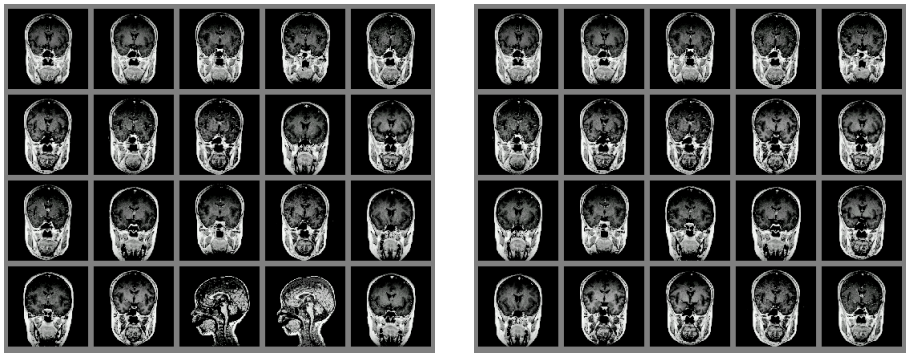


Figura B.4: P&R curves obtained for the Heterogeneous dataset.

The graph of Figure B.4 shows that the dimensionality reduction provides a significant gain in precision in the region comprising recall values under 0.4. The small recall region is the most important one, because *k*NN queries usually do not require high values of *k*. For the region of 30% of recall, the precision gains performed by the dimensionality reduction are: 1% by Relief-F; 4% by StARMiner and 5% by PreSAGe. These results testify that the dimensionality curse really influences negativelly the query results.

Figure B.5 shows the results of processing a *k*NN (k=20) query, over the query center showed at the top left from the screenshots.

Figure B.5 (a) shows the results when using the 30 original features while B.5 (b) presents the results of using the 21 features selected by PreSAGe. The major precision was obtained using the developed method (B.5 (b)), where all returned images are of the same category of the query center.



(a) (b)

Figura B.5: Results of a kNN ($k = 20$) query, over the query center showed at the top left of the screenshots. Results obtained (a) using the 30 original features; (b) using the 21 features selected by PreSAGe.

B.3 Conclusions

A new technique for feature selection is presented. The techinique employs the algorithm PreSAGe, which uses discretization of the continuous values of the feature vectors, preparing them for an association rule mining algorithm. The number of intervals generated by PreSAGe is used to determine which attributes are most relevant. The developed technique was applied to selected features of real datasets and compared to Relief-F, a well-known feature selection algorithm, achieving higher values of precision.

Apêndice C

SuGAR

The SuGAR (diagnosis **Su**ggestion **G**eneration based on **A**ssociation **R**ules) method combines low-level features automatically extracted from images with high-level knowledge given by a specialist in order to suggest a diagnosis of a new image.

C.1 The SuGAR Method

The SuGAR method is divided into: (a) the training phase and (b) the test phase. Each training image is associated with a set of keywords.

Figure C.1 shows the input and output of the SuGAR method. Features are extracted from the images, and feature vectors are used to represent the images. The feature vectors and the class label of the training images are submitted to PreSAGe, which removes irrelevant features from the feature vector and gives a discretization of the remaining features. A processed feature vector for each image is produced (see Figure C.1). The keywords of the training images and the feature vectors are used to build the transaction representation of the images. Figure C.1 provides an example of this representation. The transaction representations of all the images in the training set are submitted to the Apriori algorithm, limiting the minimum confidence to high values.

The feature vector of the test image is submitted to the HiCARe algorithm, which uses the association rules generated and suggests combinations of keywords to compose the diagnosis of

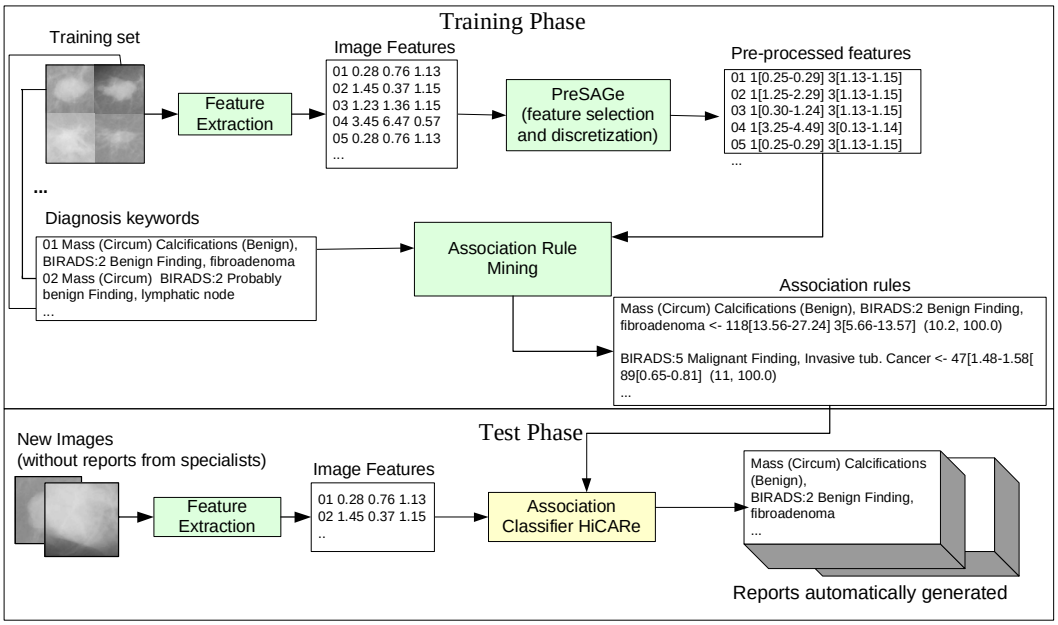


Figura C.1: Pipeline of the SuGAR method.

a test image. The process of feature extraction and the HiCARe algorithm are further detailed bellow.

C.1.1 Feature Extraction

The features of texture proposed in [Haralick et al., 1973] are extracted from images and organized into feature vectors. To perform the extraction, a *co-occurrence* matrix of the analyzed image is generated. A *co-occurrence* matrix $M(d,\theta)$ is given by the relative frequency of occurrences of two gray-level pixels i and j, separated by d pixels in the θ orientation.

First, the gray levels of the images are reduced to 16. Co-occurrence matrices are calculated for the directions of 0°, 45°, 90° and 135°, and for the distances 1, 2, 3, 4 and 5. Twenty matrices of 16x16 integer elements per image are produced. For each matrix, the seven features presented in Table C.1 are calculated, producing a feature vector of 140 elements to represent each image. This procedure is detailed in [Felipe et al., 2003].

C.1.2 HiCARe

HiCARe (Classifier based on **Hi**gh **C**onfidence **A**ssociation **R**ule Agr**ee**ments) is a new spe-

Tabela C.1: Gray-Level Texture Features and their positions in the Feature Vector.

feature	**equation**	**meaning**	**position**
Step	$\sum_i \sum_j P(i,j)$	distribution	1-20
Variance	$\sum_i \sum_j (i-j)^2 P(i,j)$	contrast	21-40
Entropy	$\sum_i \sum_j P(i,j) log(P(i,j))$	suavity	41-60
Energy	$\sum_i \sum_j P(i-j)^2$	uniformity	61-80
Homogeneity	$\sum_i \sum_j \frac{P(i-j)}{(1+\|i-j\|)}$	homogeneity	81-100
3° Moment	$\sum_i \sum_j (i-j)^3 P(i,j)$	distortion	101-120
Inv. Variance	$\sum_i \sum_j \frac{P(i,j)}{(i-j)^2}$	inv. contrast	121-140

cial classifier able to return multiple classes (keywords) when processing a test image. The following definition is required before detailing the HiCARe algorithm.

Definição C.1 *A* **match** *occurs when the image features satisfy the body part of a rule.*

The HiCARe algorithm stores all *itemsets* (set of keywords) belonging to the head of the rules in a data structure. An itemset h is returned in the suggested diagnosis *if* the condition stated in Equation C.1 is satisfied.

$$\frac{nM(h)}{nM(h)+nN(h)} \geq \beta \tag{C.1}$$

where $nM(h)$ is the number of matches of the itemset h and $nN(h)$ is the number of not-matches. A threshold $\beta\,(0 \leq \beta \leq 1)$ is employed to limit the minimal number of matches required to return an itemset in the suggested diagnosis.

C.2 Experiments

C.2.1 Experiment 1 - the "Rois" dataset

The *Rois* dataset consists of 250 images of regions of interest (encompassing lesions) taken from mammograms collected from the repository *Digital Database for Screening Mammography* - DDSM [Heath et al., 2001]. The dataset was divided in two parts: the training set that is composed of 82 images (33% of *Rois* dataset) and the test set that is composed of 168 images (67% of *Rois* dataset). The images are classified as benign or malignant, in four levels of

mammary density and in five levels of assessment (levels of the specialist's confidence in the diagnosis). Figure C.2 shows two examples of the images from the *Rois* dataset and their diagnoses.

IMAGE	DIAGNOSIS
	MALIGN, ASSESSMENT =0 DENSITY=2
	BENIGN, ASSESSMENT =3, DENSITY=2

Figura C.2: Images of the *Rois* dataset and their corresponding diagnosis.

Step A

In Step A, texture features were extracted from the images. Each image was represented by a feature vector composed of 140 features.

Step B

The image features and the class labels (malignant and benign) were submitted to the PreSAGe algorithm, using the following input parameters: $minperint = 6$, $mintofuse = 0.8$ and $valreduct = 17\%$, which are tuning parameters set by the user. PreSAGe selects 24 features as the most relevant ones, obtaining a reduction of 83% in the feature vector size.

In this step, we measure the effectiveness of PreSAGe in selecting features. For comparison purposes we also applied Relief-F [Kira & Rendell, 1992], a well-known feature selection algorithm. The 24 most relevant features returned by Relief-F were also taken to compose a feature vector. To build the Precision vs. Recall graphs, we considered three cases of feature vectors to represent the images: (a) using 140 original features, (b) using the 24 features selected by PreSAGe, (c) using the 24 features selected by Relief-F. Similarity queries were executed and the P&R graphs were constructed. Figure C.3 shows the P&R graph obtained. It shows that, even with a reduction of 83% of the feature vector size, the precision values are maintained. Moreover, PreSAGe reaches higher values of precision than Relief-F. Relief-F took 4.3 seconds to select the features and PreSAGe took 3.4 seconds (21% less time). This can represent a significant difference for larger datasets. While Relief-F executes several distance calculations, PreSAGE scans each feature value only once when performing the feature selection task.

Indeed, recall that PreSAGe performs simultaneously feature selection and discretization.

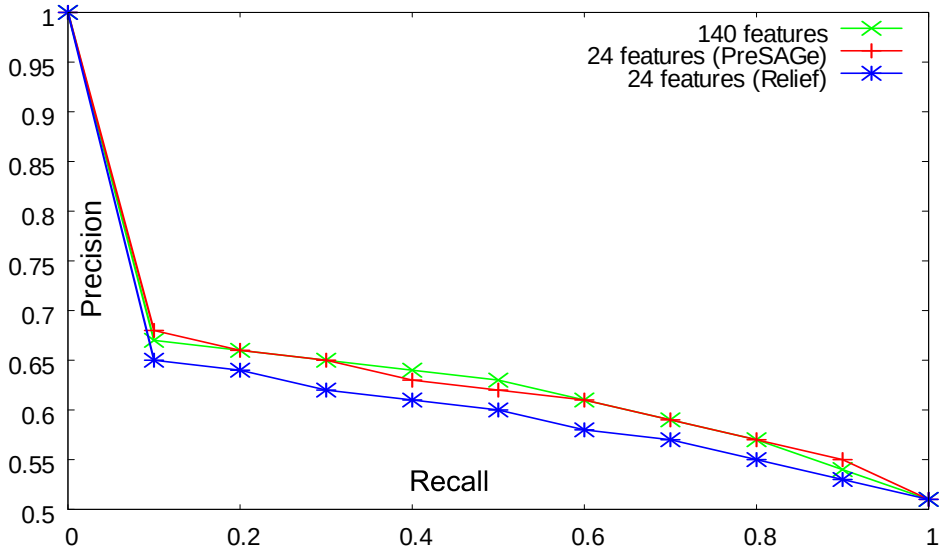


Figura C.3: P&R graph built using the *Rois* dataset represented by: 140 original features, 24 selected by PreSAGe, 24 selected by Relief-F.

Step C

The output of PreSAGe and the diagnosis information about the training images were submitted to the Apriori algorithm. The value of minimum confidence was set to 98% and the value of minimum support was set to 10%. This procedure generated 3,999 rules. in order to know how much the PreSAGe feature selection procedure reduces the number of rules, we applied the Apriori algorithm using the same configuration to the full training dataset, employing all the original 140 features. The number of rules generated was 563,693, representing an increase of around **14,000%** in the number of rules generated by Apriori. Most of the generated rules were redundant and they did not bring new knowledge. Thus, the PreSAGe feature selection procedure proved to be a critical approach in reducing the number of redundant rules and the complexity of the subsequent steps of the method. An example of a generated rule is:

102[3.95-13.24] $\rightarrow$ benign-mass (0.14, 1.0),

which says that images whose values of the 102^{nd} feature are between 3.95 and 13.24 tend to be images of benign mass. The 102^{nd} feature is the 3° Moment, which is a measure of the gray-level distortion of the image (see Table C.1). The support of the above rule is 14% and the confidence is 100%.

Step D

The images of the test set and the association rules generated in the previous step were submitted to HiCARe algorithm, using the value $\beta = 0.001$. The diagnoses suggested by the algorithm were compared to the real diagnoses of the images given by specialists and biopsy results. To validate our approach, we compared our method (considering only the diagnosis of benign and malignant mass) with two other well-known classifiers. First with C4.5, a classifier that constructs a decision tree in the training phase. Second, with Naive Bayes, a classifier that uses a probabilistic approach based on Bayes' theorem to predict the class labels. Table C.2 shows the results. Note that our method leads to higher values of sensibility, specificity and accuracy, and it also presents the smallest error rates: false positive rate = $\frac{FP}{FP+TN}$ and false negative rate =$\frac{FN}{FN+TP}$.

Tabela C.2: Results obtained by applying the developed method to *Rois* Dataset.

Measure	Developed Method	C4.5	Naive Bayes
Sensibility	95%	84%	75%
Specificity	84%	71%	63%
Accuracy	91%	79%	70%
False Positive Rate	17.6%	28,6%	37.1%
False Negative Rate	4.8%	15%	24.7%

C.2.2 Experiment 2: The Dataset "Birads"

The dataset *Birads* consists of 103 images of ROI comprising tumoral tissues, taken from mammograms collected from the Breast Imaging Reporting and Data System of the Department of Radiology of University of Vienna (www.birads.at). These ROIs are used to train new radiology students. Each image has a diagnosis composed of three main parts: Morphology, BI-RADS level and Histology.

The dataset was divided into two sets: the training set is composed of 77 images and the test set is composed of 26 images. In Step A, 140 features of texture were extracted from the images. In Step B, the features of the training images, together with the BI-RADS classification were submitted to the algorithm PreSAGe. The parameters of the PreSAGe algorithm were also set to $minperint = 6$, $mintofuse = 0.8$ and $valreduct = 17$. The algorithm selected 24 features

as relevant. In Step C, the output from Step B and the diagnosis information about the training images are submitted to the algorithm Apriori. The value of minimum confidence was set to 100% and the value of minimum support was set to 8%. The association rule algorithm found 19 rules in Step C. In Step D, the features of test images were submitted to HiCARe using $\beta = 0.005$. Table C.3 shows the values of accuracy obtained by applying our method, taking into consideration the three main parts of a diagnosis (morphology, BI-RADS and histology).

Tabela C.3: Values of Accuracy obtained by applying the developed method to *Birads* dataset.

Distinguishing	Accuracy
Morphology	92%
BI-RADS value	84%
Histology	84%

The highest accuracy value (92%) was achieved in identifying morphological features (mass and calcification). Most mistakes in finding the proper BI-RADS values occured suggesting a value adjacent to the real one. For example, the algorithm suggests 5 and the specialist suggests 4. Discerning between two adjacent BIRADS values can be difficult even to a veteran radiologist, because of the high similarity of the images between consecutive levels. Hence, 84% of accurracy in stating the BI-RADS value is a very good result. Figure C.4 shows a test image and its diagnosis (given by a specialist) and the diagnosis suggested by our developed method.

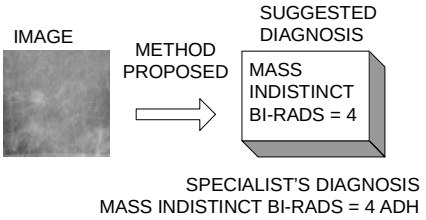


Figura C.4: Example of a result of the developed method.

C.3 Conclusions

The increasing use of image exams in the last 25 years has greatly contributed to improve the diagnosing of diseases as well as to enhance the health care of patients. However, the volume of images has grown at a fast pace and the specialists have been unable to keep up with

diagnosing. The resultant bottleneck is precisely what computer aided diagnosis (CAD) systems are built to overcome. We presented the SuGAR method, which is based on association rules, to generate suggestions for diagnosis of medical images, and which can be integrated into a CAD. Our approach is divided into four main steps: feature extraction; discretization and feature selection; association rule mining; generation of diagnosis suggestions. The second step uses a new algorithm, called PreSAGe, which executes two tasks in a single step: feature selection and discretization. The feature selection process speeds up and reduces the complexity of the whole method, making it faster than traditional approaches. The last step is performed by the HiCARe algorithm, which generates a suggested diagnosis by assigning multiple keywords to a test image. The results applied to real databases show that the developed method achieves high sensitivity (up to 95%) and accuracy (up to 92%) in the task of supporting decision making.