Mineração de imagens médicas utilizando características de forma

Alceu Ferraz Costa

Mineração de imagens médicas utilizando características de forma[1]

Alceu Ferraz Costa

Orientadora: **Profa. Dra. Agma Juci Machado Traina**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

USP – São Carlos
Junho de 2012

[1] Trabalho realizado com apoio financeiro da FAPESP - Processo 2009/12905-2

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Ferraz Costa, Alceu
F837m Mineração de imagens médicas utilizando características de forma / Alceu Ferraz Costa; orientadora Agma Juci Machado Traina. -- São Carlos, 2012.
96 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012.

1. Mineração de Imagens. 2. Diagnóstico Auxiliado por Computador. 3. Extração de Características. 4. Classificação de Imagens. 5. Imagens Médicas. I. Juci Machado Traina, Agma, orient. II. Título.

Agradecimentos

À minha orientadora, Profa. Dra. Agma J. M. Traina, que pela orientação, apoio e ensinamentos me apresentou à pesquisa e tornou possível esse trabalho. Muito obrigado!

Ao meu pai e à minha mãe, que sempre apoiaram e incentivaram meus estudos. Esta dissertação é resultado de toda atenção e carinho que sempre recebi.

À minha namorada Lourdes pelo amor e carinho. *Gracias por estar a mi lado, gracias por todo.*

Aos meus amigos, do ICMC e do GBdI que compartilharam muitos momentos alegres que certamente não irei esquecer. Agradecimentos especiais ao Gabriel, Glauco e Lúcio que ajudaram com a revisão desta dissertação.

Aos meus irmãos, Lucas e Tiago.

À FAPESP e CNPq pelo apoio financeiro.

À Deus.

Resumo

Bases de imagens armazenadas em sistemas computacionais da área médica correspondem a uma valiosa fonte de conhecimento. Assim, a mineração de imagens pode ser aplicada para extrair conhecimento destas bases com o propósito de apoiar o diagnóstico auxiliado por computador (*Computer Aided Diagnosis - CAD*). Sistemas *CAD* apoiados por mineração de imagens tipicamente realizam a extração de características visuais relevantes das imagens. Essas características são organizadas na forma de vetores de características que representam as imagens e são utilizados como entrada para classificadores. Devido ao problema conhecido como lacuna semântica, que corresponde à diferença entre a percepção da imagem pelo especialista médico e suas características automaticamente extraídas, um aspecto desafiador do *CAD* é a obtenção de um conjunto de características que seja capaz de representar de maneira sucinta e eficiente o conteúdo visual de imagens médicas. Foi desenvolvido neste trabalho o extrator de características *FFS* (*Fast Fractal Stack*) que realiza a extração de características de forma, que é um atributo visual que aproxima a semântica esperada pelo ser humano. Adicionalmente, foi desenvolvido o algoritmo de classificação *Concept*, que emprega mineração de regras de associação para predizer a classe de uma imagem. O aspecto inovador do *Concept* refere-se ao algoritmo de obtenção de representações de imagens, denominado *MFS-Map* (*Multi Feature Space Map*) e também desenvolvido neste trabalho. O *MFS-Map* realiza agrupamento de dados em diferentes espaços de características para melhor aproveitar as características extraídas no processo de classificação. Os experimentos realizados para imagens de tomografia pulmonar e mamografias indicam que tanto o *FFS* como a abordagem de representação adotada pelo *Concept* podem contribuir para o aprimoramento de sistemas *CAD*.

Abstract

Medical image databases represent a valuable source of data from which potential knowledge can be extracted. Image mining can be applied to knowledge discover from these data in order to help CAD (Computer Aided Diagnosis) systems. The typical set-up of a CAD system consists in the extraction of relevant visual features in the form of image feature vectors that are used as input to a classifier. Due to the semantic gap problem, which corresponds to the difference between the humans' image perception and the features automatically extracted from the image, a challenging aspect of CAD is to obtain a set of features that is able to succinctly and efficiently represent the visual contents of medical images. To deal with this problem it was developed in this work a new feature extraction method entitled *Fast Fractal Stack* (*FFS*). *FFS* extracts shape features from objects and structures, which is a visual attribute that approximates the semantics expected by humans. Additionally, it was developed the *Concept* classification method, which employs association rules mining to the task of image class prediction. The innovative aspect of *Concept* refers to its image representation algorithm termed *MFS-Map* (*Multi Feature Space Map*). *MFS-Map* employs clustering in different feature spaces to maximize features' usefulness in the classification process. Experiments performed employing computed tomography and mammography images indicate that both *FFS* and *Concept* methods for image representation can contribute to the improvement of CAD systems.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Símbolos

$\mathcal{I}$	Conjunto de itens que podem ocorrer em uma base de dados transacional.
$\mathcal{T}$	Conjunto de transações de uma base de dados transacional.
I	Imagem.
$\mathbf{I}$	Conjunto de imagens.
W	Palavras chaves associadas a uma imagem.
$\mathbf{W}$	Conjunto de palavras chaves associadas a várias imagens.
$\mathcal{W}$	Conjunto os valores de palavras chaves.
$\vec{v}$	Vetor de características.
$\mathbf{V}$	Conjunto de vetores de características.
$\hat{V}$	Vetor de características transacional.
$\hat{\mathbf{V}}$	Conjunto de vetores de características transacionais.
$\mathbf{S}$	Conjunto de regras de associação.
I_b	Imagem binária.
Δ	Imagem de contornos.
$\mathcal{F}$	Espaço de características.
ε	Extrator de características.
E	Conjunto de extratores de características.
$\mathcal{D}$	Dimensão fractal.
$\mathcal{D}_0$	Dimensão fractal de Haussdorf.
M'	Função de mapeamento para o algoritmo *MFS-Map*.
g	Centróide de agrupamento.

G	Conjunto de centróides de agrupamentos.
θ	Agrupamento.
Θ	Conjunto de agrupamentos.

Capítulo 1

Introdução

1.1 Motivação

A computação tem apoiado o desenvolvimento da medicina em diversas áreas: em sistemas de apoio a coleta de dados clínicos e exames por imagens, na modelagem de objetos e estruturas anatômicas, no desenvolvimento de simuladores de procedimentos, na organização das informações obtidas, entre outras. Os sistemas computacionais fornecem versatilidade ao processo de armazenamento e transmissão de exames médicos digitalizados e, como resultado de sua operação, um grande volume de dados médicos são gerados, processados e armazenados.

Diversas modalidades de imagens médicas, tais como ultrassom, raio-X, ressonância magnética e tomografia computadorizada fazem parte dos dados médicos armazenados nestes sistemas computacionais. As informações contidas nas imagens são complementadas por laudos compostos de textos. Desta maneira, esse grande volume de dados é uma valiosa fonte de conhecimento que pode ser utilizada para o auxílio ao diagnóstico médico e para o ensino da medicina. Assim, é grande a importância do desenvolvimento de técnicas que permitam a descoberta de conhecimento em bases de imagens médicas para apoiar o médico em sua tarefa diária de tomada de decisões, aumentado a precisão, confiabilidade e eficiência dos diagnósticos elaborados pelo especialista. Esse apoio computacional pode atuar como uma junta médica virtual, ao trazer para o especialista o conhecimento armazenado em exames e diagnósticos relacionados.

Para se realizar a descoberta de conhecimento em bases de imagens médicas é importante cruzar as representações de características visuais de baixo nível das imagens, obtidas por meio de técnicas de visão computacional [Datta 08], com as informações de alto nível provenientes dos laudos associados às imagens. No entanto, devido ao problema conhecido como lacuna semântica, esta tarefa não é trivial. A lacuna semântica [Deserno 09] refere-se à dificuldade em obter das características visuais de baixo nível informações que correspondam à interpretação que o especialista médico tem da imagem. Por este motivo, é crucial o desenvolvimento de pesquisas em técnicas de extração de características que reduzam a lacuna semântica, bem como técnicas de mineração de dados para analisar as informações contidas em tais características para a descoberta de conhecimento em bases de imagens médicas.

1.2 Definição do Problema e Objetivos

Uma das maiores dificuldades enfrentadas no desenvolvimento de sistemas de diagnóstico auxiliado por computador (*Computer Aided Diagnosis - CAD*) apoiados por técnicas de mineração de imagens é o problema de inconsistência entre a representação de baixo nível e a interpretação de alto nível das imagens. Para amenizar este problema, o projetista de um sistema *CAD* apoiado por mineração de imagens deve recorrer a características de baixo nível que possuam o máximo de correlação possível com as informações semânticas da imagem. Este projeto de Mestrado considerou como hipótese que a forma de objetos presentes na imagem é um atributo visual que aproxima à semântica esperada pelo ser humano. Por este motivo, o trabalho realizado teve como um dos focos o desenvolvimento de técnicas de extração de características de forma voltadas para o domínio de imagens médicas.

O segundo foco deste trabalho consistiu no desenvolvimento de técnicas de mineração de imagens para aprimorar o sistema *IDEA*. O *IDEA* [Ribeiro 08] é um sistema *CAD* que surgiu do trabalho conjunto entre o Grupo de Base de Dados e Imagens (GBdI) e do Centro de Ciências de Imagens e Física Médica (CCIFM), ambos da USP, em incorporar aos sistemas de arquivamento e comunicação de imagens (*Picture Archive and Communication System - PACS*) funcionalidades de auxílio ao diagnóstico.

1.3 Principais Contribuições deste Projeto de Mestrado

Para realizar o aprimoramento do sistema *IDEA*, foi desenvolvido o extrator de características *FFS* (*Fast Fractal Stack*) que emprega análise fractal para medir a complexidade de contornos de estruturas e objetos presentes em uma imagem. Os resultados obtidos com o *FFS* para a tarefa de classificação de doenças pulmonares difusas em imagens de tomografia do pulmão foram publicadas no *ACM Workshop on Medical Multimedia Analysis and Retrieval (MMAR 2011)*[Costa 11], junto ao *ACM Multimedia 2011*.

A segunda contribuição deste trabalho consistiu na realização de atividades que, inicialmente, tiveram como foco o aprimoramento do método *IDEA*. No entanto, com o desenvolvimento de tais atividades um novo algoritmo de classificação de imagens, denominado *Concept*, foi proposto. O *Concept* e o *IDEA* compartilham a estratégia de usar mineração de regras de associação para classificar imagens, mas apresentam diferenças significativas. A principal dessas diferenças está no modo como as representações das imagens na forma de *itemsets* são obtidas. As representações na forma de *itemsets* são necessárias para mineração de regras de associação e, no método *IDEA*, sua obtenção é realizada por meio da extração de características juntamente com a discretização de atributos. Ou seja, cada item que compõe um *itemset* representa um intervalo de discretização de um determinado atributo numérico. No método *Concept* é utilizado um novo algoritmo desenvolvido durante o projeto de Mestrado, o *MFS-Map* (*Multi Feature Space Map*), para extrair representações na forma de *itemsets* das imagens. A grande vantagem do *MFS-Map* está no fato de que os itens representam regiões de diferentes espaços de características. Em tais regiões as imagens são visualmente similares e, desta maneira, os itens obtidos pelo *MFS-Map* carregam informações semânticas valiosas para o processo de classificação.

1.4 Organização da Dissertação

Esta dissertação está organizada da seguinte maneira. No capítulo 2 é discutida a aplicação de técnicas de mineração de imagens no contexto de diagnóstico auxiliado por computador. No capítulo 3, são definidos e apresentados os conceitos relacionados à extração de características. O capítulo 4 descreve o método de extração *FFS* proposto neste projeto de mestrado. O método *Concept* para classificação de imagens é descrito no capítulo 5. As conclusões, contribuições e possibilidades de pesquisas futuras decorrentes deste trabalho são apresentadas no capítulo 6.

Parte I

Conceitos

Capítulo 2

Mineração de Imagens Aplicada ao Diagnóstico Auxiliado por Computador

A crescente disparidade existente entre o volume de imagens geradas devido ao avanço das tecnologias de aquisição e armazenagem e a habilidade de humanos analisarem tais dados não é um fenômeno recente. Esse fato é confrontado com a necessidade de analisar e extrair conhecimentos úteis de dados armazenados em sistemas computacionais. Em [Burl 99] é realizada uma análise das técnicas então existentes que aliam a mineração de dados ao processamento de imagens com o objetivo de realizar a mineração de imagens de maneira automática.

De maneira mais abrangente, a mineração de imagens pode ser caracterizada como uma disciplina que lida com a extração de conhecimento, padrões e relações não explicitamente armazenados em imagens e dados alfanuméricos associados [Hsu 02]. Trata-se ainda de um campo interdisciplinar, que faz uso da visão computacional, processamento de imagens, recuperação de imagens, mineração de dados, aprendizado de máquina, bancos de dados e inteligência artificial [Rui 07, Wang 09, Becker 10].

Ainda que a mineração de dados seja um aspecto de grande importância no processo de extração de conhecimentos e padrões em bases de imagens, a tarefa de mineração de imagens não pode ser vista como simplesmente uma aplicação da mineração de dados a um domínio específico, pois na mineração de dados tradicional os dados estão representados na forma tabular, relacional ou de grafos. Na mineração de imagens existe o desafio de extrair representações significativas das imagens uma vez que os valores de pixels individuais de imagens não têm significado semântico. A semântica é obtida pela análise de vizinhança de pixels, podendo a mesma estar relacionada com o contexto de aplicação.

Para lidar com os desafios apresentados, o processo de mineração de imagens pode ser dividido em quatro etapas: pré-processamento de imagem, extração de características, integração e mineração, conforme ilustrado no diagrama da figura 2.1. A etapa de pré-processamento tem como objetivo atenuar ruídos e outras características visuais indesejadas ao mesmo tempo em que realça características importantes para aplicação. A etapa de extração de características, discutida em maiores detalhes no capítulo 3, tem a finalidade de gerar uma representação das características visuais de baixo nível da imagem

para o processo de mineração [Datta 08]. Na etapa de integração, a representação de imagem obtida é associada a dados textuais que descrevem as imagens. Por fim, na etapa de mineração, são aplicados algoritmos de mineração de dados (os quais podem ser adaptados ao domínio de imagens) para se extrair conhecimentos da base de dados.

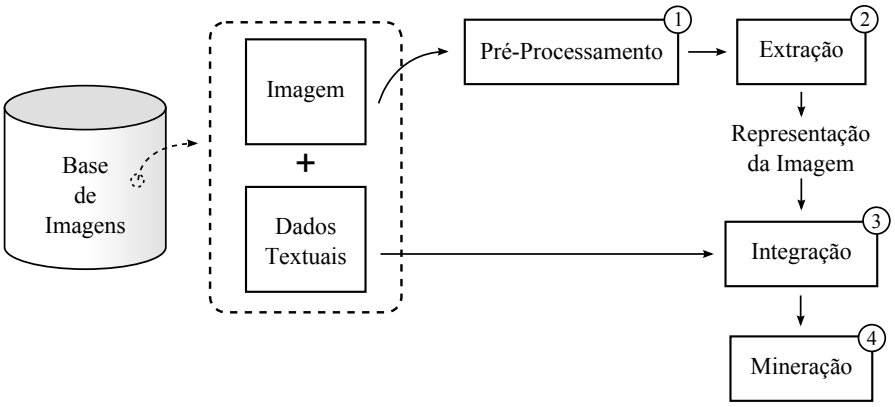


Figura 2.1: As etapas do processo de mineração de imagens.

De acordo com [Hsu 02], as pesquisas no campo de mineração de imagens seguem duas vertentes: a de domínio específico e a de propósito geral. Considerando-se as etapas do processo de mineração de imagens apresentadas no diagrama da figura 2.1, a vertente de domínio específico tem como foco as etapas de pré-processamento e extração, uma vez que seu principal objetivo é desenvolver técnicas que obtenham representações visuais das imagens que sejam mais significativas quanto possível [Datta 08]. As pesquisas de propósito geral, por sua vez, procuram entender a interação existente entre as características visuais de baixo nível das imagens e a percepção de alto nível que os seres humanos têm da mesma [Bugatti 09, Silva 09]. Neste trabalho foram utilizadas abordagens que integram as duas vertentes mencionadas, aplicando-as à tarefa de extração de informações e padrões de bases de imagens médicas para o desenvolvimento de sistemas de auxílio ao diagnóstico baseado em imagens.

2.1 Mineração de Dados

Conforme ilustrado no diagrama da figura 2.1, uma vez obtida uma representação adequada da imagem e realizada sua integração com os dados textuais associados, a próxima etapa do processo de mineração de imagens consiste na realização da tarefa de mineração propriamente dita. A mineração de dados é o processo de extração de conhecimento de grandes conjuntos de dados [Han 05]. Segundo [Pang-Ning 05], as tarefas de mineração de dados podem ser categorizadas como sendo preditivas ou descritivas. As tarefas preditivas têm como objetivo prever o valor de um determinado atributo alvo com base no valor dos demais atributos. Já as tarefas descritivas, procuram extrair padrões tais como correlações, tendências e agrupamentos que podem ser utilizados para descrever os dados sendo analisados.

Algumas das principais tarefas de mineração de dados são:

Classificação: Trata-se de uma tarefa de predição utilizada para atributos alvo do tipo discreto. A classificação pode ser utilizada, por exemplo, para se prever o diagnóstico de um paciente.

Regressão: Assim como a classificação, a regressão é uma tarefa preditiva. No entanto, o atributo alvo é do tipo contínuo como, por exemplo, na previsão da temperatura em um determinado local e dia.

Análise de Agrupamentos: O objetivo desta tarefa é encontrar grupos de objetos de modo que aqueles que pertençam a um mesmo grupo são mais similares entre si que objetos que pertençam a grupos diferentes [Jain 10]. A análise de agrupamentos pode ser utilizada, por exemplo, para encontrar grupos de clientes que apresentem comportamentos similares.

Detecção de Anomalias: É uma tarefa que consiste em encontrar observações com características significativamente diferente das demais presentes no conjunto de dados. Essas observações são denominadas anomalias (*outliers*) e sua identificação pode ser aplicada, por exemplo, no problema de detecção de fraudes em instituições financeiras [Deriche 93].

Outra importante tarefa de mineração de dados é a análise de associações. Seu objetivo é encontrar e analisar padrões em uma base de dados que apresentem uma forte associação entre variáveis e seus valores. Uma vez que neste projeto de mestrado foi explorada a mineração de imagens por meio de regras de associação, a seção que se segue tem como objetivo discutir em maiores detalhes a tarefa de análise de associações. Adicionalmente, na seção 2.2 é realizada uma breve descrição da tarefa de análise de agrupamentos devido à sua aplicação no desenvolvimento deste trabalho.

2.1.1 Análise de Associações

Empresas de grande porte normalmente produzem grandes volumes de dados operacionais. Uma rede de supermercados, por exemplo, coleta diariamente dados na forma de transações que correspondem a itens de uma compra realizada em suas lojas. A tabela 2.1 ilustra a organização deste tipo de dado. Cada linha corresponde a uma transação que possui um identificador denominado *TID* (*Transaction Identifier*) e um conjunto de itens comprado por um cliente. A tarefa de análise de associações foi proposta em 1993 [Agrawal 93] com o objetivo de se encontrar regras de associações, que definam relações existentes entre itens de uma grande base de dados. Por exemplo, a partir das transações da tabela 2.1, a regra de associação $\{\text{Fraldas}\} \Rightarrow \{\text{Cerveja}\}$ poderia ser extraída, uma vez que muitos dos clientes que compram fralda também compram cerveja.

Tabela 2.1: Exemplo de dados organizados na forma de transações.

TID	Transação
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Refrigerante}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Refrigerante}

Um conceito importante na análise de associações é o de *itemset* (conjunto de itens). Sendo $\mathcal{I} = \{i_1, i_2, \ldots, i_d\}$ o conjunto de todos os itens que podem ocorrer em uma transação, um *itemset* é definido como um conjunto de zero ou mais itens de $\mathcal{I}$, sendo que um *itemset* que possua k itens é denominado *k-itemset*. Por exemplo, o *itemset* {Pão, Leite} é um 2-*itemset*, pois contém dois itens. Sendo $\mathcal{T} =$

$\{t_1, t_2, \ldots, t_N\}$ o conjunto de todas as transações da base de dados, define-se o suporte de um *itemset* X como sendo o número de transações $t_j \in \mathcal{T}$ que contêm X. Formalmente, o suporte de um *itemset* X pode ser expresso da seguinte maneira [Pang-Ning 05]:

$$\sup(X) = |\{t_i | X \subseteq t_i,\ t_i \in \mathcal{T}\}| \tag{2.1}$$

sendo que $|\cdot|$ denota o número de elementos do conjunto. No caso da tabela 2.1, o *itemset* {Pão, Leite} tem suporte igual a três, uma vez que três das cinco transações da tabela contêm todos os dois itens.

Considerando os conceitos apresentados, uma regra de associação é uma expressão de implicação na forma $X \Rightarrow Y$, onde X e Y são dois *itemsets* disjuntos e X e Y são chamados, respectivamente, de antecedente e conseqüente da regra. Para se determinar o quão forte é uma regra, são utilizadas as medidas de suporte e confiança. O suporte mede o quão freqüente é uma determinada regra no conjunto de dados, enquanto que a confiança mede a freqüência com que os itens em Y ocorrem em transações que contêm X. Sendo N o total de transações existentes na base de dados, as medidas de suporte e confiança podem ser definidas da seguinte maneira:

$$\text{Suporte:}\quad \sup(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{N} \tag{2.2}$$

$$\text{Confiança:}\quad \text{conf}(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)} \tag{2.3}$$

Tomando como exemplo a tabela 2.1 e a regra $\{\text{Fraldas}\} \Rightarrow \{\text{Cerveja}\}$, temos que $X = \{\text{Fraldas}\}$, $Y = \{\text{Cerveja}\}$ e $X \cup Y = \{\text{Fraldas}, \text{Cerveja}\}$. Uma vez que o número de transações na tabela 2.1 é cinco o suporte da regra é dado por:

$$\sup(\{\text{Fraldas}\} \Rightarrow \{\text{Cerveja}\}) = \frac{\sup(\{\text{Fraldas}, \text{Cerveja}\})}{N} = \frac{3}{5} \tag{2.4}$$

A medida de confiança da regra é calculada como:

$$\text{conf}(\{\text{Fraldas}\} \Rightarrow \{\text{Cerveja}\}) = \frac{\sup(\{\text{Fraldas}, \text{Cerveja}\})}{\sup(\{\text{Fraldas}\})} = \frac{3}{4} \tag{2.5}$$

O problema de minerar regras de associação consiste em se encontrar todas as regras fortes que ocorrem em uma base de dados, ou seja, dado um conjunto $\mathcal{T}$ de transações, deseja-se encontrar todas as regras com suporte maior que *minSup* (denominadas regras freqüentes) e com confiança maior que *minConf*, onde *minSup* e *minConf* são valores de limiar mínimos para as métricas de suporte e confiança definidos conforme a aplicação ou pelo usuário. Um algoritmo de força bruta para resolver o problema consiste em gerar todas a regras possíveis e então computar os valores de suporte e confiança para cada uma delas, descartando aquelas que não atenderem às restrições dos limiares *minSup* e *minConf*. No entanto, essa abordagem não é viável, uma vez que o total de regras que podem ser geradas cresce exponencialmente considerando o número de itens existentes na base de dados.

Notando que o suporte de uma regra $X \Rightarrow Y$ depende somente do suporte do *itemset* $X \cup Y$ (equação 2.2), se o *itemset* $X \cup Y$ é infreqüente então todas as regras que poderiam ser geradas a partir de $X \cup Y$ também serão infreqüentes. Desta maneira, uma primeira abordagem para tornar o processo de mineração de regras de associação mais eficiente consiste em dividir o problema em duas etapas. Na primeira são descartados os *itemsets* infreqüentes e na segunda etapa apenas os *itemsets* restantes são utilizados

para gerar regras que atendam a restrição da confiança mínima.

O método utilizado para gerar os *itemsets* freqüentes tem um papel fundamental no desempenho do algoritmo. Supondo que em uma base de dados existam d itens, então o total de *itemsets* que podem ser gerados, excluindo o *itemset* vazio, é $2^d - 1$. Uma vez que o total de *itemsets* cresce exponencialmente com relação ao número de itens na base de dados, computar o suporte para cada um desses *itemsets* torna-se inviável em muitas aplicações práticas. O algoritmo *Apriori* [Agrawal 94], discutido na seção que se segue, apresenta uma solução mais eficiente para o problema de geração de *itemsets* freqüentes.

2.1.2 O Algoritmo *Apriori*

Uma importante propriedade apresentada pela medida de suporte é a anti-monoticidade: o valor de suporte de um *itemset* nunca excede o valor de suporte de seus subconjuntos. Sendo $J = 2^{\mathcal{I}}$ o conjunto potência de $\mathcal{I}$ (conjunto de todos os subconjuntos de $\mathcal{I}$), pode-se dizer que a medida de suporte é anti-monotônica pois:

$$\forall X, Y \in J : (X \subseteq Y) \implies \sup(X) \geq \sup(Y) \tag{2.6}$$

A anti-monoticidade do suporte é também conhecida como princípio apriori. Esse princípio consiste no fato de que, se um *itemset* é freqüente, então todos os seus subconjuntos também o são. O algoritmo *Apriori*, proposto em [Agrawal 94], faz uso deste princípio para controlar a quantidade de *itemsets* candidatos gerados. *Itemsets* que possuam subconjuntos sabidamente infreqüentes podem ser descartados, dispensando o cálculo da medida de suporte. Tomando como exemplo o látice de *itemsets* da figura 2.2, se o *itemset* {ab} não atender o critério do suporte mínimo, então, pelo princípio apriori, pode-se descartar os *itemsets* {abc}, {abd} e {abcd}, devido a estes terem {ab} como subconjunto.

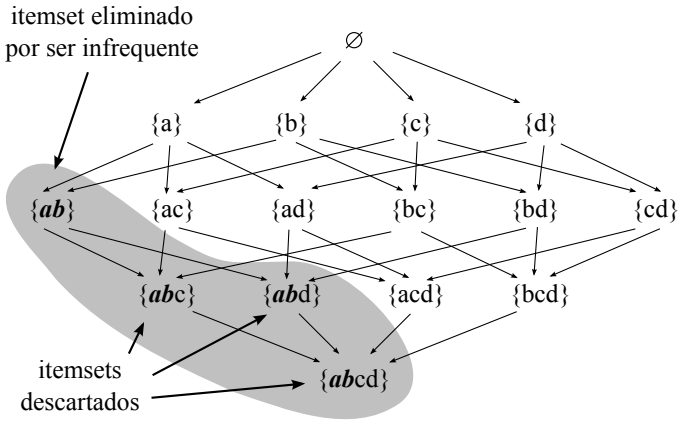


Figura 2.2: O princípio apriori pode ser utilizado para descartar *itemsets* compostos por subconjuntos infreqüentes. No exemplo, {ab} é infreqüente e, por este motivo, os *itemsets* {abc}, {abd} e {abcd} podem ser descartados, por também serem considerados infreqüentes.

O algoritmo para geração de *itemsets* freqüentes do algoritmo *Apriori* consiste em, inicialmente, encontrar todos os 1-*itemsets* e seus respectivos valores de suporte. É então, inicializada a variável k com valor 1 e as três etapas que se seguem são repetidas até que nenhum *itemset* seja identificado:

Geração de Itemsets Candidatos: A partir dos k-*itemsets*, são gerados os $(k+1)$-*itemsets* candidatos.

Descarte: Os $(k+1)$-*itemsets* candidatos que tenham como subconjunto *k-itemsets* infreqüentes são descartados.

Eliminação: É computado o valor de suporte para os $(k+1)$-*itemsets* candidatos restantes, eliminando aqueles que não atendam o critério do suporte mínimo. O valor da variável k é incrementado em uma unidade.

Para gerar os *itemsets* candidatos, o algoritmo *Apriori* utiliza o método $F_{k-1} \times F_{k-1}$, onde F_{k-1} denota o conjunto de todos os $(k-1)$-*itemsets* freqüentes. O método consiste em manter ordenados os itens dentro de um *itemset* e então agrupar dois $(k-1)$-*itemsets* somente se os seus $k-2$ itens forem idênticos. Ou seja, sendo $X = \{x_1, x_2, \ldots, x_{(k-1)}\}$ e $Y = \{y_1, y_2, \ldots, y_{(k-1)}\}$ dois $(k-1)$-*itemsets* freqüentes, o algoritmo $F_{k-1} \times F_{k-1}$ realiza seu agrupamento somente se a condição que se segue é satisfeita:

$$x_i = y_i \text{ para } i = 1, 2, \ldots, k-2 \text{ e } x_{k-1} \neq y_{k-1} \tag{2.7}$$

É importante notar que método $F_{k-1} \times F_{k-1}$ pode gerar *itemsets* candidatos infreqüentes. Por este motivo, realiza-se em seguida a etapa de descarte dos *itemsets* candidatos que tenham subconjuntos infreqüentes. Por fim, na etapa de eliminação, é feita a contagem do suporte dos *itemsets* candidatos não descartados, eliminando aqueles que não atendam ao critério do suporte mínimo.

A geração de candidatos freqüentes pelo algoritmo *Apriori* é descrita no algoritmo 2.1. Nas linhas 1-2 são encontrados os 1-*itemsets* e seus respectivos valores de suporte e inicializada a variável k. A geração de *itemsets* candidatos juntamente com o descarte daqueles que contêm subconjuntos infreqüentes ocorre na linha 5. A atualização do valor de suporte dos *itemsets* candidatos é realizada nas linhas 6-11. Ao invés de se comparar cada transação t com todos os *itemsets* candidatos, é feita na linha 7 uma enumeração dos *itemsets* candidatos contidos em t que são atribuídos à variável C_t. A enumeração pode ser feita de maneira eficiente se os itens de cada transação t forem armazenados de maneira ordenada. Nas linhas 8-10 os *itemsets* candidatos enumerados têm seu valor de suporte atualizado. Por fim, na linha 12, os *itemsets* candidatos que não atendam o critério do suporte mínimo são eliminados.

Algoritmo 2.1 Geração de candidatos freqüentes do algoritmo *Apriori*.

Entrada: Conjunto de itens $\mathcal{I} = \{i_1, i_2, \ldots, i_d\}$ e conjunto de transações $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$.
Saída: Conjunto de *itemsets* freqüentes $\bigcup F_k$.

1: $k \leftarrow 1$
2: $F_k \leftarrow \{i \mid i \in \mathcal{I} \wedge \sup(\{i\}) \geq N \times minSup\}$ // *Atribui a F_k todos os 1-*itemsets *freqüentes.*
3: **repetir**
4: $\quad k \leftarrow k+1$
5: $\quad C_k \leftarrow$ **geraItemsetsCandidatos**(F_{k-1}) // *$F_{k-1} \times F_{k-1}$ seguido por descarte.*
6: $\quad$ **para** cada transação $t \in \mathcal{T}$ **faça**
7: $\quad\quad C_t \leftarrow \{c \mid c \subseteq t \wedge c \subseteq C_k\}$ // *C_t são os candidatos que pertencem a t.*
8: $\quad\quad$ **para** cada *itemset* candidato $c \in C_t$ **faça**
9: $\quad\quad\quad \sup(c) \leftarrow \sup(c) + 1$
10: $\quad\quad$ **fim para**
11: $\quad$ **fim para**
12: $\quad F_k \leftarrow \{c \mid c \in C_k \wedge \sup(c) \geq N \times minSup\}$ // *Atribui a F_k todos os k-*itemsets *freqüentes.*
13: **até que** $F_k = \emptyset$
14: **retorna** $\bigcup F_k$

A próxima etapa do algoritmo *Apriori* consiste em gerar regras a partir dos *itemsets* freqüentes encontrados. É possível extrair uma regra de associação realizando o particionamento de um *itemset* freqüente Y em dois conjuntos não vazios, X e $Y - X$, tal que a regra $X \Rightarrow Y - X$ tenha confiança maior que $minConf$. Apesar da medida de confiança não satisfazer a propriedade de monoticidade, se uma regra $X \Rightarrow Y - X$ não satisfizer o critério da confiança mínima, então qualquer regra $X' \Rightarrow Y - X'$, onde X' é um subconjunto de X, não satisfará o limiar mínimo de confiança. Essa propriedade permite realizar o descarte de possíveis regras candidatas, tornando a etapa de geração de regras mais eficiente.

2.1.3 Classificadores Associativos

A classificação associativa é uma abordagem em mineração de dados na qual associações fortes entre padrões de características e classes são mineradas para construir classificadores, denominados de classificadores associativos. Desta maneira, a classificação associativa pode ser vista como a integração entre duas tarefas de mineração de dados: a de classificação e a de mineração de regras de associações. Esta integração é realizada por meio de um subconjunto especial de regras de associação nas quais o conseqüente está restrito ao atributo classe. Na literatura, tais regras são denominadas de regras de associação de classificação (*Classification Association Rules - CARs*).

O problema de se construir um classificador associativo pode ser dividido em duas etapas principais: (i) minerar as *CARs* e (ii) determinar a classe de um dado objeto de teste por meio das regras geradas. Uma vez que uma *CAR* está na forma $X \Rightarrow \{c'\}$, onde $X = \{x_1, x_2, \cdots, x_n\}$ representa um conjunto de itens derivados dos atributos do conjunto de dados e c' um valor para o atributo classe, uma abordagem que pode ser empregada para mineração consiste em encontrar todos os *itemsets* no formato $\{x_1, x_2, \cdots, x_n, c'\}$. Para este propósito, pode-se empregar algoritmos de mineração de *itemsets* freqüentes tradicionais, como o *Apriori*. De fato, esta abordagem é amplamente adotada por métodos de indução de classificadores associativos [Liu 98, Ribeiro 08].

Quando o número de atributos do conjunto de dados é muito alto, sobretudo quando o limiar mínimo de suporte *minSup* é baixo, o custo computacional do algoritmo Apriori na tarefa de mineração de *CARs*, em termos de tempo de *CPU* e uso de memória pode ser extremamente alto. Para tornar o processo de mineração de *CARs* mais eficiente, algoritmos como o L^3 [Baralis 02] empregam o método *FP-Growth*. O método *FP-Growth* (*Frequent Pattern Growth*), proposto em [Han 00], minera regras de associação por meio de uma árvore de prefixos denominada de *FP-Tree*. Na *FP-Tree* cada *itemset* é representado por um caminho na árvore, sendo que dois *itemsets* diferentes podem compartilhar parte de um dado caminho. Esta propriedade da árvore permite que o conjunto de dados seja representado utilizando menos memória. Adicionalmente, a construção da *FP-Tree* exige que o conjunto de dados seja percorrido somente duas vezes, reduzindo o número de leituras a disco.

Existem ainda métodos de treinamento de classificadores associativos que não empregam algoritmos tradicionais de regras de associação. Um exemplo é o algortimo *CPAR* (*Classification based on Predictive Associations Rules*)[Han 03] que adota uma estratégia gulosa na qual a construção de uma regra consiste em crescer o antecedente da regra um item por vez, sendo que o item a ser adicionado à regra é aquele que maximiza a confiança da regra. A principal desvantagem do método de mineração do algoritmo *CPAR* é seu alto tempo de execução quando comparado com os demais métodos [Thabtah 07].

Outro aspecto relevante a respeito dos classificadores asssociativos é o conceito de ordenação entre regras. Por exemplo, em [Liu 98], as regras mineradas são ordenadas com respeito à sua métrica

de confiança, sendo que o suporte e a ordem de geração da regra são empregados com critérios de desempate. Essa ordenação é utilizada pelo algoritmo tanto para descartar regras desnecessárias como também para classificar um dado objeto de teste.

Uma vez obtido o conjunto de regras final, existem duas abordagens principais para se determinar a classe de um objeto de teste. A primeira delas, empregada em [Liu 98, Baralis 02], envolve definir uma ordem entre as regras geradas com base nas medidas de suporte e confiança das regras. A classe predita é aquela que apresenta a maior precedência dentre todas as que são aplicáveis ao objeto de teste. Uma segunda estratégia, adotada em [Han 03, Antonie 04, Ribeiro 08], consiste em encontrar um valor de escore para cada uma das classes a partir do conjunto de regras que são aplicáveis ao objeto sendo classificado. A classe que apresentar o maior valor de escore é aquela que será retornada pelo classificador.

A tabela 2.2 apresenta um resumo dos principais aspectos de diferentes métodos de classificação associativa. Na primeira coluna, (*Mineração de Regras*), é apresentado o algoritmo de mineração de *CARs* empregado por cada método. Na coluna *ordenação de regras* são apresentados os critérios, por ordem de precedência, empregados para ordenar as regras mineradas. Caso ocorra empate para um determinado critério, o seguinte é utilizado para realizar o desempate. Por fim, a coluna *Modelo de Predição* corresponde à estratégia empregada para classificar um *itemset*. Adicionalmente, por ter sido utilizado como base para o desenvolvimento deste projeto de mestrado, o método *IDEA* é descrito em maiores detalhes na seção 2.3.

Tabela 2.2: Comparação entre métodos de classificação associativa.

Nome	**Mineração de Regras**	**Ordenação de Regras**	**Modelo de Predição**	**Referência**
CBA	*Apriori*	1. Confiança; 2. Suporte; 3. Ordem de geração da regra.	Única regra	[Liu 98]
L^3	*FP-Growth*	1. Confiança; 2. Suporte; 3. Número de itens na regra; 4. Ordem lexicográfica do antecedente da regra.	Única regra	[Baralis 02]
CPAR	Gulosa	1. Confiança; 2. Suporte; 3. Número de itens na regra.	Múltiplas regras	[Han 03]
ARC-PAN	*Apriori*	1. Confiança.	Múltiplas regras	[Antonie 04]
IDEA	*Apriori*	Não utiliza ordenação de regras.	Múltiplas regras	[Ribeiro 08]

Outro aspecto da classificação associativa a ser discutido se refere ao formato do conjunto de dados. Conforme discutido na seção 2.1.1, para se realizar a tarefa de análise de associações os dados devem estar no formato transacional, ou seja, ser representados por um conjunto de *itemsets* $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$,

onde t_i corresponde a um *itemset*. Um conjunto de dados sobre o qual se aplica a tarefa de classificação, no entanto, está no formato tabular, ou seja, corresponde a uma tabela composta por $m+1$ colunas. As linhas desta tabela são denominadas de tuplas e as colunas correspondem aos m atributos $\{A_1, A_2, \cdots, A_m\}$ e aos valores de classe **C**, sendo que os atributos podem ser tanto categóricos (assumem valores restritos a um conjunto finito) ou numéricos. Desta maneira, para minerar as *CARs*, é necessário mapear os dados do formato tabular para o formato transacional. Para os atributos categóricos basta mapear cada um de seus valores para um item. Para os atributos numéricos, a abordagem adotada pelos métodos existentes na literatura consiste em realizar a discretização e mapear cada um dos intervalos de discretização para um item.

2.2 Análise de Agrupamentos

O objetivo da tarefa de análise de agrupamentos é dividir um conjunto objetos em agrupamentos de modo que objetos em um mesmo grupo sejam mais similares que objetos de grupos diferentes. No contexto de mineração de imagens, é usual que cada objeto corresponda a uma representação extraída da imagem, conforme apresentado no diagrama da figura 2.1.

Assumindo que cada imagem seja representada por um vetor de características $\vec{v} = (v_1, v_2, \cdots, v_n)$, ou seja, um conjunto de n atributos numéricos que corresponde a um ponto em um espaço n-dimensional, um algoritmo clássico de análise de agrupamento que pode ser aplicado é o *k-Means* [Pang-Ning 05]. No algoritmo *k-Means* cada agrupamento possui um centróide e os vetores de características são associados ao agrupamento que apresenta o centróide mais próximo.

A inicialização do algoritmo consiste em definir k centróides aleatórios, onde k é um parâmetro de entrada do algoritmo. Uma vez definidos os centróides iniciais, o algoritmo consiste em duas etapas principais:

1. Associar cada vetor de cacterísticas ao agrupamento que possua o centróide mais próximo;
2. Recalcular as posições dos centróides. A posição do centróide será o centro de massa do conjunto de vetores de características associados ao respectivo agrupamento.

O algoritmo é finalizado quando as posições dos centróides permanecem fixas após uma iteração. Um dos problemas apresentado pelo algoritmo *k-Means* é que o número k de agrupamentos deve ser fornecido como parâmetro de entrada, pois uma escolha inadequada para o parâmetro k pode trazer resultados insatisfatórios. Uma possível abordagem para resolver este problema consiste em executar o *k-Means* para vários valores de k e escolher o resultado avaliado como mais adequado.

Em [Pelleg 00] é proposto o algoritmo *x-Means* no qual essa abordagem é aprimorada. Depois de encontrado o conjunto inicial de agrupamentos o algoritmo não é reiniciado para testar um novo valor para o parâmetro k. Ao invés disso, os agrupamentos são avaliados por meio de um escore baseado no critério de informação bayesiano (*Bayesian Information Criterion - BIC*). Cada um dos centróides é então dividido em dois novos centróides, que são obtidos movendo o centróide original em direções opostas ao longo de um vetor aleatório no espaço de características. Cada par de centróides resultantes é avaliados pelo escore baseado no *BIC*. Caso o escore *BIC* seja superior ao do centróide original, o par de centróides é mantido e o número k é incrementado em uma unidade. Caso contrário, os par de centróides é descartado e o centróide original é mantido.

2.3 O Método *IDEA*

Esta seção tem como objetivo discutir o método *IDEA* [Ribeiro 08], que foi utilizado como base para o desenvolvimento deste projeto de mestrado. *IDEA* é um acrônimo para "*Image Diagnosis Enhancement through Association rules*" e consiste em um método que emprega regras de associação para auxiliar no diagnóstico de imagens médicas. O *IDEA* apresenta como saída do processo de mineração de imagens um conjunto de palavras chaves ranqueadas por uma medida de convicção. Essas palavras chaves são utilizadas como uma fonte de informações para auxiliar o especialista médico no processo de diagnóstico. Adicionalmente, dependendo do tipo de dado textual associado às imagens, o método permite que múltiplas hipóteses de diagnóstico sejam sugeridas para uma mesma imagem.

O método *IDEA* faz uso de dois algoritmos também introduzidos em [Ribeiro 08]: o *Omega* e o *ACE*. O algoritmo *Omega* realiza a discretização e seleção de atributos, possibilitando a conversão da representação das imagens para o formato transacional, utilizado por algoritmos de mineração de regras de associação como o *Apriori*. O *ACE* (*Associative Classifier Engine*), por sua vez, é um classificador associativo que faz uso de regras de associação para sugerir as palavras chaves que irão compor o diagnóstico associado com valores de convicção para cada sugestão. A discussão sobre o Método *IDEA* será feita em duas etapas. Na primeira, será discutido o processo de mineração de regras de associação a partir das imagens de treinamento. Na segunda etapa, é discutido como as regras mineradas são utilizadas pelo classificador associativo *ACE* para sugerir palavras chaves para as imagens a serem diagnosticadas.

2.3.1 Mineração de Regras de Associação

Para realizar a mineração de regras de associação, o método *IDEA* toma como entrada o conjunto de todas as imagens de treinamento $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ e o conjunto de palavras chaves associadas às imagens $\mathbf{W} = \{W_1, W_2, \ldots, W_N\}$. Caso a imagem I_i seja uma mamografia, então as palavras chaves W_i associadas a I_i poderiam ser, por exemplo, {massa, oval, maligno}. A partir de **I** e **W**, é retornado um conjunto de regras de associações **S** que será utilizado pelo classificador associativo *ACE* para sugerir as palavras chaves que irão compor o diagnóstico de uma imagem. O diagrama da figura 2.3 ilustra as quatro etapas do processo de mineração de regras de associação no método *IDEA*.

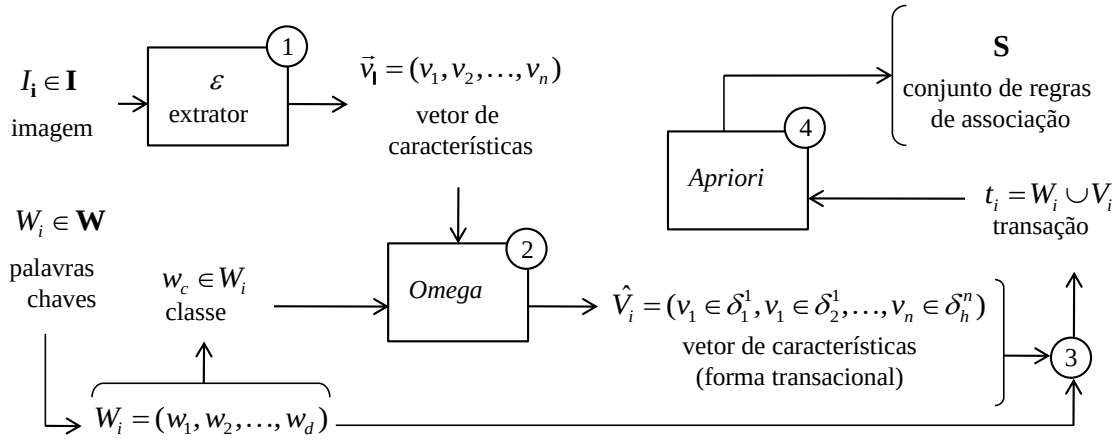


Figura 2.3: Mineração de regras de associação pelo método *IDEA*.

A primeira etapa (figura 2.3) consiste em extrair uma representação sucinta da imagem. A notação $\varepsilon(\cdot)$ é utilizada para representar a função de extração, que sumariza as informações visuais da imagem I na forma de um vetor de características $\vec{v} = (v_1, v_2, \ldots, v_n)$. No contexto de mineração de dados, esse vetor organiza um conjunto de n atributos. Uma vez que o algoritmo *Apriori* trabalha somente com dados no formato transacional, a segunda etapa do método consiste em converter a vetor de características em um *itemset*. Para tanto, é utilizado o algoritmo *Omega*, que realiza simultaneamente seleção e discretização de atributos. Uma característica importante deste método é que ele precisa tomar como entrada a classe de cada uma das imagens de treinamento. O algoritmo proposto em [Ribeiro 08] utiliza um subconjunto de palavras chaves, escolhidas por um especialista, como as classes possíveis de diagnóstico. No caso de uma mamografia, por exemplo, as palavras chaves *benigno* e *maligno* poderiam ser as classes possíveis.

Para realizar o processo de discretização, o *Omega* faz uso de uma medida de inconsistência local, que mede quanto dos elementos de um intervalo de discretização não pertencem à classe majoritária. Sempre que dois intervalos de discretização contíguos são agrupados, obtém-se um novo intervalo para o qual a medida de inconsistência é maior ou igual a dos intervalos originais. O processo de discretização é realizado agrupando-se intervalos contíguos até que seja alcançado um limiar máximo de inconsistência. Para selecionar atributos, é definida a métrica de inconsistência global, que consiste em uma média da medida de incosistência local de cada um dos intervalos, ponderada pelo número de elementos em cada um deles. Atributos que apresentem um valor de inconsistência global maior que um limiar máximo são descartados.

A saída do algoritmo é um *itemset* $\hat{V}$ cujos itens estão na forma $v_i \in \delta_k^i$, onde δ_k^i representa o k-ésimo intervalo de discretização do atributo v_i. Ou seja, cada item de $\hat{V}$ indica em qual intervalo de discretização se encontra cada um dos valores do vetor de características original. A figura 2.4 exemplifica como é feita a conversão do vetor de características para um *itemset*. Cada um dos intervalos de discretização δ_k^i resulta em atributo binário que toma o valor 1 se v_i está contido em δ_k^i. No exemplo, v_1 está contido em δ_2^1 pois $\delta_2^1 = [0.3, 1.0]$ e v_2 está contido em δ_2^2 pois $\delta_2^2 = [0.1, 0.4)$. Por fim, $\hat{V}$ (também denominado vetor de características transacional) é composto pelos itens que correspondem ao atributos binários que assumiram o valor 1. Na figura 2.4, $\vec{v}^T$ denota a transposição do vetor $\vec{v}$.

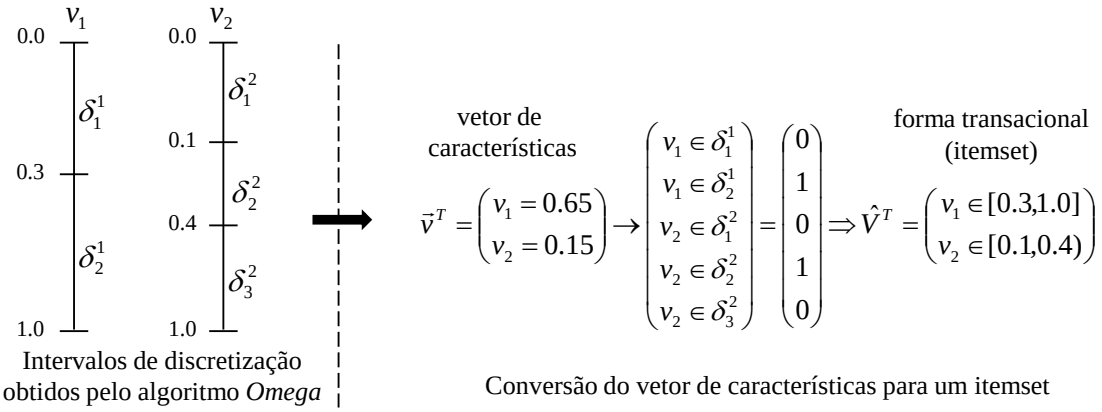


Figura 2.4: Conversão de um vetor de características para um *itemset* pelo algoritmo *Omega*.

Para realizar a mineração de regras de associação, o vetor de características na forma transacional $\hat{V}$ é agrupado com as palavras chaves associadas à imagem, compondo uma transação (etapa três do diagrama

da figura 2.3). O conjunto de todas as transações obtidas é então submetido ao algoritmo *Apriori* (seção 2.1.2), adotando-se duas restrições às regras que podem ser geradas: o atencedente é composto somente por itens do vetor de características transacional e o conseqüente somente por palavras chaves. A equação 2.8 exemplifica uma possível regra de associação minerada de um conjunto de imagens de mamografias.

$$\{v_3 \in [0.5, 0.7], v_5 \in [0.2, 0.4]\} \Rightarrow \{\text{maligno}, \text{calcificação}\} \tag{2.8}$$

A regra da equação 2.8 pode ser interpretada da seguinte maneira: imagens que apresentem os atributos v_3 e v_5 do vetor de características contidos, respectivamente, nos intervalos $[0.5, 0.7]$ e $[0.2, 0.4]$ tendem a ser diagnosticadas com as palavras chaves *maligno* e *calcificação*. Uma última observação a ser feita sobre a mineração de regras de associação diz respeito ao valor da medida confiança utilizado. Nos experimentos descritos em [Ribeiro 08], os autores afirmam que é importante utilizar um valor alto de confiança (maior que 97%) para que não ocorra a degradação do desempenho do classificador associativo. O algoritmo 2.2 descreve o processo de mineração de regras de associação do método *IDEA*.

Algoritmo 2.2 Mineração de regras de associação pelo método *IDEA*.

Entrada: Conjunto de imagens de treinamento $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ e conjunto de palavras chaves de cada imagem $\mathbf{W} = \{W_1, W_2, \ldots, W_N\}$.
Saída: Conjunto de regras de associação **S**.

1: **para** cada imagem $I \in \mathbf{I}$ **faça**
2: $\quad \vec{v} \leftarrow \varepsilon(I)$ *// Extração do vetor de características.*
3: $\quad \mathbf{V} \leftarrow \mathbf{V} \cup \{\vec{v}\}$ *// Adiciona o vetor de características a* $\boldsymbol{V}$.
4: **fim para**
5: $\hat{\mathbf{V}} \leftarrow \mathbf{Omega}(\mathbf{V}, \mathbf{W})$ *// Converte os vetores de característica para o formato transacional.*
6: **para** $i \in \{1..N\}$ **faça**
7: $\quad t_i \leftarrow \{W_i \cup \hat{V}_i | W_i \in \mathbf{W} \wedge \hat{V}_i \in \hat{\mathbf{V}}\}$
8: $\quad T \leftarrow T \cup \{t_i\}$ *// Adiciona transação* t_i *a* T.
9: **fim para**
10: $\mathbf{S} \leftarrow \mathbf{Apriori}(T)$
11: **retorna S**

As linhas 2 e 3 correspondem à primeira etapa do diagrama da figura 2.3, na qual é feita a extração dos vetores de características. A linha 5 corresponde à etapa dois, na qual é realizada a conversão dos vetores de características para forma transacional por meio do algoritmo *Omega*. Nas linhas 6-9 as palavras chaves são agrupadas com o vetor de características transacional de cada imagem para gerar transações (etapa 3 do diagrama da figura 2.3). Por fim, na linha 10 é realizada a mineração de regras de associação utilizando o algoritmo *Apriori*.

2.3.2 O Algoritmo *ACE*

O classificador associativo *ACE*, responsável por sugerir as palavras chaves W_S que irão compor o diagnóstico, toma como entrada a imagem a ser diagnosticada e o conjunto de regras de associação **S** obtido na primeira etapa do método *IDEA*. Cada uma das palavras chaves em W_S tem um valor de convicção associado. Quanto maior o valor da medida de convicção de uma palavra chave, maior é a probabilidade de que ela de fato componha o diagnóstico.

A medida de convicção usa o conceito de casamento entre uma regra de associação e um vetor de características. Um vetor de características *satisfaz* uma regra quando seus atributos satisfazem todos os itens presentes no antecedente da regra. Um vetor de característica *satisfaz parcialmente* uma regra quando seus atributos satisfazem ao menos um mas não todos os itens do antecedente da regra. A equação 2.9 ilustra esses conceitos:

$$\vec{v}^T = \begin{pmatrix} v_1 = 0.6 \\ v_2 = 0.2 \\ v_3 = 0.4 \end{pmatrix} \begin{array}{c} \xrightarrow{\text{satisfaz parcialmente}} \\ \xrightarrow{\text{não satisfaz}} \\ \xrightarrow{\text{satisfaz}} \end{array} \begin{cases} \{v_1 \in [0.5, 0.7], v_2 \in [0.8, 0.9]\} \Rightarrow \text{benigno} \\ \{v_2 \in [0.8, 0.9], v_3 \in [0.1, 0.2]\} \Rightarrow \text{benigno} \\ \{v_3 \in [0.3, 0.7]\} \Rightarrow \text{benigno} \end{cases} \tag{2.9}$$

A função $\mathrm{M_T}(w)$ é utilizada para contar quantas vezes a palavra chave w fez parte de uma regra de associação que *satisfez* um vetor de características. As funções $\mathrm{M_P}(w)$ e $\bar{\mathrm{M}}(w)$, por sua vez, contam quantas vezes a palavra chave *satisfez parcialmente* e *não satisfez* o vetor de características. $\mathrm{M_T}(w)$, $\mathrm{M_P}(w)$ e $\bar{\mathrm{M}}(w)$ são utilizadas para calcular a medida de convicção a partir da seguinte expressão:

$$\mathrm{conv}(w) = \frac{3\mathrm{M_T}(w) + \mathrm{M_P}(w)}{3\mathrm{M_T}(w) + \mathrm{M_P}(w) + \bar{\mathrm{M}}(w)} \tag{2.10}$$

Tomando como exemplo o vetor de características da equação 2.9 tem-se que $\mathrm{M_T}(\text{benigno}) = \mathrm{M_P}(\text{benigno}) = \bar{\mathrm{M}}(\text{benigno}) = 1$. Desta maneira, o valor da medida de convicção para a palavra chave benigno é:

$$\mathrm{conv}(w = \text{benigno}) = \frac{3 \cdot 1 + 1}{3 \cdot 1 + 1 + 1} = \frac{4}{5} \tag{2.11}$$

Adicionalmente, palavras chaves w que apresentem $\mathrm{M_T}(w) < 1$ ou $\mathrm{conv}(w) < minConv$, onde *minConv* é um limiar mínimo para o valor da medida de convicção, são descartadas. O algoritmo 2.3 descreve o classificador associativo *ACE*.

Na linha 1, o vetor de características da imagem a ser diagnosticada é extraído. Nas linhas 2-12 são computados os valores de $\mathrm{M_T}(w)$, $\mathrm{M_P}(w)$ e $\bar{\mathrm{M}}(w)$ para todas as palavras chaves $w \in \mathcal{W}$, onde $\mathcal{W}$ denota o conjunto de todas as palavras chaves existentes na base de dados. Com base nos valores obtidos, é calculado o valor da medida de convicção para cada uma das palavras chaves. Caso a palavra chave atenda à condição expressa na linha 14, ela é adicionada ao conjunto W_S de palavras chaves sugeridas.

2.4 Diagnóstico Auxiliado por Computador

O conceito de se criar um hospital que não dependa de imagens em filme ("*filmless*") por meio da adoção de sistemas de arquivamento e comunicação de imagens (*Picture Archive and Communication System - PACS*) data da década de 1980 [Inamura 95]. A motivação para se realizar a transmissão e arquivamento de imagens de maneira digital surgiu das limitações da radiografia baseada em filmes [Bick 99]. Uma radiografia, por exemplo, só pode estar localizada em um local em um dado momento, seu transporte é uma tarefa que consome muito tempo e seu armazenamento demanda uma considerável quantidade de espaço físico, como ilustrado na figura 2.5.

A tarefa de um *PACS* consiste na aquisição digital de diversas modalidades de imagens médicas (tais como ultrassom, raio-X e ressonância magnética), armazenamento destas imagens em uma base de

Algoritmo 2.3 Sugestão de palavras chaves por meio do algoritmo *ACE*.

Entrada: Imagem I a ser diagnosticada e conjunto de regras de associação $\mathbf{S}$.
Saída: Conjunto de palavras chaves sugeridas W_S com valor de convicção $\text{conv}(\cdot)$ associado.

1: $\vec{v} \leftarrow \varepsilon(I)$ // *Extração do vetor de características.*
2: **para** cada regra $\hat{V} \Rightarrow W \in \mathbf{S}$ **faça**
3: **para** cada item $w \in W$ **faça**
4: **se** $\hat{V}$ satisfaz $\vec{v}$ **então**
5: $\mathrm{M_T}(w) \leftarrow \mathrm{M_T}(w) + 1$
6: **senão se** $\hat{V}$ satisfaz parcialmente $\vec{v}$ **então**
7: $\mathrm{M_P}(w) \leftarrow \mathrm{M_P}(w) + 1$
8: **senão**
9: $\bar{\mathrm{M}}(w) \leftarrow \bar{\mathrm{M}}(w) + 1$ // *$\hat{V}$ não satisfaz $\vec{v}$.*
10: **fim se**
11: **fim para**
12: **fim para**
13: **para** cada $w \in \mathcal{W}$ **faça**
14: **se** $\mathrm{M_T}(w) \geq 1 \wedge \text{conv}(w) = \dfrac{3\mathrm{M_T}(w) + \mathrm{M_P}(w)}{3\mathrm{M_T}(w) + \mathrm{M_P}(w) + \bar{\mathrm{M}}(w)} \geq minConv$ **então**
15: $W_S \leftarrow W_S \cup \{w\}$ // *Adiciona w ao conjunto de palavras chaves sugeridas.*
16: **fim se**
17: **fim para**
18: **retorna** W_S

(a) (b)

Figura 2.5: Arquivamento de imagens médicas em filme no Departamento de Radiologia Clínica da Universidade de Muenster [Bick 99]. (a) Biblioteca onde são armazenados os filmes recentemente obtidos. Após um período de seis meses, os filmes são movidos para uma biblioteca secundária. (b) Técnica em radiologia realizando o transporte de filmes radiológicos.

dados central e em sua disponibilização mediante, por exemplo, a uma requisição de um especialista médico [Marques 09]. Estatísticas levantadas por [van De Wetering 09] em 2006 apontaram que a taxa de adoção média de *PACS* entre países europeus é de 33% e que deve-se esperar um aumento da adoção de *PACS* nos próximos anos.

Desta maneira, os dados contidos nos *PACS*, que incluem imagens médicas e laudos textuais não estruturados, representam uma fonte crescente de informações médicas valiosas. No entanto, em virtude da complexidade e volume dos dados a serem analisados, os profissionais da área da saúde ainda não se beneficiam de grande parte dessa fonte de conhecimento. Para se resolver esse problema, a mineração

de imagens pode ser utilizada para extrair informações que possibilitem o desenvolvimento de técnicas de auxílio ao diagnóstico médico.

O uso de computadores na análise de imagens médicas com o propósito de se realizar diagnóstico médico, conforme levantamento feito em [Giger 08], teve seu início na década de 1950, motivado pelas dificuldades enfrentadas pelos radiologistas na detecção e caracterização de anormalidades sutis presentes no exame. Isto se deve, dentre outros fatores, às limitações do sistema visual humano, distrações e fadiga do especialista. As primeiras abordagens tinham como objetivo realizar um diagnóstico automatizado, ou seja, sem a intervenção do radiologista. No entanto, de acordo com [Doi 07], estas tentativas não obtiveram sucesso, sobretudo devido a uma expectativa excessivamente alta que se tinha da tecnologia computacional na época, uma vez que as técnicas de processamento de imagens ainda se encontravam em um estado inicial e pelo fato do poder computacional ainda ser muito limitado.

Na década de 1980 uma nova abordagem foi proposta na qual o resultado da análise feita pelo computador é utilizado pelo radiologista ao invés de substituí-lo no processo de decisão. Essa abordagem ficou conhecida como Diagnóstico Auxiliado por Computador ou *CAD* (*Computer Aided Diagnosis*). O *CAD* pode ser definido como o diagnóstico feito pelo especialista médico apoiado no resultado da análise computacional dos dados. Desta maneira, a utilização do *CAD* pode ser vista como uma segunda opinião no processo de decisão do especialista, tendo como objetivo reduzir erros de interpretação e a variação do diagnóstico entre um mesmo radiologista e entre diferentes radiologistas.

Diferentemente do diagnóstico automatizado, o *CAD* não tem como objetivo alcançar um desempenho superior ao do especialista mas sim prover informações que complementem a análise feita inicialmente, aumentando sua confiabilidade. Caso um sistema de diagnóstico automático apresente um desempenho inferior ao de um radiologista, torna-se extremamente difícil justificar seu uso, uma vez que pacientes não aceitariam um diagnóstico de qualidade inferior àquele que poderia ser obtido por um especialista [Doi 07]. No entanto, se o mesmo método for utilizado como um sistema *CAD*, como mostrado no diagrama da figura 2.6, é possível alcançar um desempenho superior àquele obtido unicamente pelo radiologista. Isso se deve ao fato de informações complementares terem sido introduzidas no processo de decisão do radiologista.

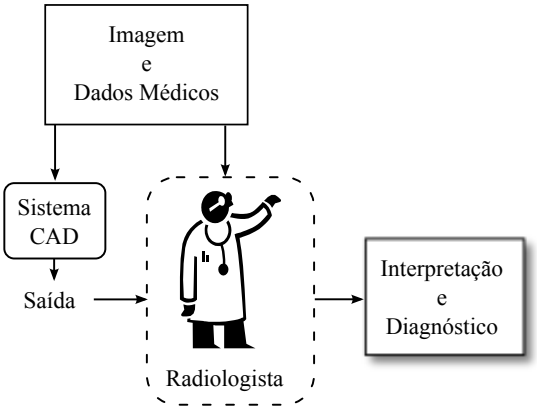


Figura 2.6: Utilização de um sistema *CAD* no processo de interpretação e diagnóstico de imagens médicas.

A seção que se segue tem como objetivo realizar um levantamento de trabalhos na área de diagnóstico auxiliado por computador apoiado por técnicas de mineração de imagens. Em seguida, na seção 2.6, é

realizada uma breve discussão sobre a avaliação de desemepenho de sistemas de diagnóstico auxiliado por computador.

2.5 Exemplos de Aplicações de Mineração de Imagens em Diagnóstico Auxiliado por Computador

Existem diversas abordagens para se realizar o diagnóstico auxiliado por computador. Em [Naqa 05], por exemplo, para o problema de detecção de microcalcificações em mamografias são listadas quatro categorias distintas de métodos *CAD*: 1) métodos baseados em realce de imagens; 2) métodos de modelagem estocástica; 3) métodos de decomposição multiescala; 4) métodos baseados em aprendizado de máquina. Por terem sido o foco deste projeto de Mestrado, nesta seção são apresentados trabalhos correlatos que se apoiam na mineração de imagens para realizar o diagnóstico auxiliado por computador. Ou seja, a análise computacional do exame a ser diagnosticado é realizada com base no conhecimento extraído de bases de imagens médicas previamente laudadas.

Em [Wang 04] foram aplicadas técnicas de mineração de regras de associação para classificação de imagens de mamografia. Para tanto, foi utilizada uma base com 1745 imagens divididas em casos com diagnóstico maligno, benigno e normal. Note que um caso pode, e em geral contém mais de uma imagem. De cada massa tumoral segmentada por um especialista foram extraídas três medidas de forma, que correspondem a três atributos contínuos. O algoritmo de mineração de regras de associação utilizado é o *Apriori* [Agrawal 94], que necessita que os atributos sejam discretos, ou seja, estejam na forma de um *itemset*. Por este motivo, cada uma das três medidas de forma, que assumem valores entre 0.0 e 1.0, são discretizadas em 10 intervalos de tamanho 0.1. As regras geradas são restritas a somente aquelas que apresentem a classe (benigna ou maligna) em seu conseqüente. Para realizar a classificação de uma nova massa tumoral a partir das regras mineradas, as três medidas de forma são extraídas e então são selecionadas as regras que sejam aplicáveis aos valores obtidos. O rótulo atribuído é o da regra forte com maior valor para a métrica de confiança. O método apresenta duas limitações importantes: 1) a segmentação das massas tumorais não é feita de maneira automática, exigindo que um especialista realize essa tarefa, o que pode ser inviável considerando grandes volumes de imagens; 2) a discretização das medidas de forma é feita utilizando intervalos de tamanho fixo, o que pode resultar em perda de informação. Uma solução para o segundo problema é apresentada na seção 2.3 utilizando o algoritmo *Omega*.

A esclerose múltipla (EM) é uma doença neurológica auto-imune crônica, que resulta em múltiplas áreas de desmielinização [1] inflamatória no sistema nervoso central. Em [Loizou 11] é proposto um método de classificação com o propósito de diferenciar lesões de EM que irão resultar em casos leves e moderados da doença. Para tanto, foram selecionados 38 pacientes com sintomas leves da doença e lesões detectáveis por imagens de ressonância magnética. Para cada paciente foi realizada uma ressonância magnética e as lesões foram identificadas e segmentadas por um neurologista especialista em EM. Após 6-12 meses do exame de ressonância, foi realizada por um neurologista uma avaliação dos sintomas da doença e os casos foram divididos em dois grupos: os que permaneceram com sintomas leves e os que desenvolveram sintomas moderados da doença. Foram então extraídos, de cada umas das lesões, descritores de textura propostos pelos autores baseados na representação por modulação de freqüência

[1] A mielina é uma substância dielétrica na matéria branca que isola as terminações dos nervos.

e amplitude (FM e AM) multiescala da imagem. Esses descritores foram utilizados para treinar um classificador *SVM* (*Support Vector Machine*). Segundo os autores, os resultados obtidos mostram que o método proposto pode diferenciar, com uma acurácia de até 86%, lesões que irão resultar em casos com sintomas leves e lesões que irão resultar em casos com sintomas moderados da doença.

A baciloscopia, que consiste na análise por microscópio de lâminas com escarro de pacientes, é, segundo [Khutlang 10], responsável pela maior parte das detecções de casos de tuberculose (TB) atualmente. Em [Khutlang 10] é proposto um método para detecção de bacilos da tuberculose em lâminas coradas pelo método Ziehl-Neelsen (ZN). Para tanto, é realizada inicialmente a segmentação de objetos que são candidatos a serem bacilos por meio da classificação dos pixels como sendo fundo ou bacilos. Foi utilizado um classificador bayesiano treinado com imagens segmentadas por especialistas. Dos objetos segmentados, foi extraído o contorno paramétrico (ver seção 3.2 para maiores detalhes) a partir do qual foram computados descritores de forma e cor. Uma vez que os descritores de forma extraídos resultam em um vetor de características com muitos componentes, foi realizada seleção de atributos. Os autores realizaram experimentos com três diferentes métodos de seleção de atributos: CFS (*Correlation-based feature selection*) [Hall 00], SFFS/SBFS (*Sequential floating foward/backward selection*) e *branch and bound*. Para classificação final foi utilizado um conjunto de treinamento com 6901 objetos, sendo 72,4% rotulados como bacilos e os 27,6% restantes como não-bacilos. Utilizando validação cruzada com dez partições, foram avaliados os classificadores *k*NN, rede neural RBF de uma camada e classificadores *SVM*. Para a rede neural, que obteve os melhores resultados, foram atingidos os valores de 98,53%, 97,71% e 99,13% para as métricas de acurácia, sensitividade e especifidade, respectivamente.

2.6 Avaliação de Sistemas de Diagnóstico Auxiliado por Computador

A avaliação de desempenho é um aspecto de grande importância na pesquisa de sistemas *CAD*. Em [Chan 90] foi realizado um teste clínico para a tarefa de detecção de microcalcificações em mamografias. Um primeiro grupo de radiologistas realizou a tarefa com auxílio computacional e um segundo grupo realizou a tarefa sem o auxílio. Foi constatado um aumento significativo na taxa de detecção do primeiro grupo e os resultados foram essenciais para demonstrar a viabilidade de sistemas *CAD* [Giger 08]. No entanto, segundo [Heath 00a], poucos algoritmos foram testados em ambiente clínico devido, sobretudo, aos custos associados a experimentos desta natureza.

Uma alternativa menos custosa para se avaliar um método *CAD* consiste em se utilizar bases de imagens previamente diagnosticadas. Um exemplo é a base de imagens de mamografias disponibilizada publicamente pela *University of South Florida*. A DDSM (*Digital Database for Screening Mammography*) [Heath 00b] atualmente contém 2620 casos. Cada caso possui as quatro imagens de um exame de mamografia: as projeções craniocaudal e médio-lateral oblíqua de cada mama. Casos que apresentem anormalidades tais como massas tumorais ou calcificações, incluem um contorno da mesma bem como palavras chaves para sua descrição.

Desta maneira, no contexto de *CAD*, o objetivo de uma avaliação de desempenho que emprega bases de imagens é verificar se o método considerado é capaz de fornecer um diagnóstico compatível com aquele presente no laudo do exame. Uma medida básica que pode ser utilizada para se comparar dois métodos *CAD* é a acurácia, que consiste na proporção dos casos de teste que foram diagnosticados corretamente. Sua utilidade, no entanto, é limitada em situações nas quais existe um desbalanceamento dos possíveis valores de diagnóstico. Por exemplo, se em um conjunto de teste de mamografias 1% dos

casos são malignos, então um método *CAD* que classifique todos o casos como benignos pode atingir uma acurácia de 99% ainda que ele falhe completamente em detectar casos malignos.Para contornar essa limitação da acurácia, pode-se utilizar uma série de medidas que são derivadas da matriz de confusão binária.

A matriz de confusão sumariza o número de casos diagnosticados de modo correto ou incorreto por um método *CAD*. Para se construir a matriz é utilizado o conceito de classe positiva *P* e classe negativa *N*. Em geral, as classes positivas e negativas referem-se, respectivamente, ao diagnóstico que ocorre com menor a maior freqüência [Pang-Ning 05]. No contexto de mamografias, por exemplo, o diagnóstico maligno poderia ser considerado a classe positiva e o diagnóstico benigno a classe negativa. A tabela 2.3 ilustra o conceito de matriz de confusão. Nela, as entradas *VP* e *VN* correspondem, respectivamente, ao número de casos classificados corretamente como positivos e negativos. As entradas *FP* e *FN*, por sua vez, correspondem, respectivamente, ao número de casos classificados incorretamente como positivos e negativos.

Tabela 2.3: Matriz de confusão binária.

		Classe Predita	
		P	**N**
Classe Verdadeira	**P**	*VP*	*FN*
	N	*FP*	*VN*

A tabela 2.4 sumariza as principais medidas que podem ser derivadas da matriz de confusão binária. Precisão e revocação são usadas, sobretudo em aplicações onde a detecção da classe positiva é considerada mais importante que a detecção da classe negativa. Em geral, ao se aumentar a precisão ou revocação de um determinado método *CAD* observa-se uma redução no valor da outra medida. Ou seja, existe uma relação de compromisso ("*tradeoff*") entre as duas métricas. A medida-F_1 pode ser utilizada para sumarizar a precisão e a revocação em uma única métrica por meio de uma média harmônica.

2.6.1 Curvas *ROC*

A análise de curvas *ROC* (*Receiver Operating Characteristic*) é um método para avaliação, organização e seleção de sistemas de diagnósticos [Prati 08]. Trata-se de uma alternativa à avaliação por meio de medidas que ilustra graficamente o *tradeoff* entre a TVP (taxa de verdadeiros positivos) e a TFP (taxa de falsos positivos). Na curva *ROC*, a TVP corresponde ao eixo y e a TFP ao eixo x. Um determinado método *CAD* que apresente valores binários para o diagnóstico é representado por um ponto no espaço *ROC* que é obtido calculando sua TVP e TFP a partir de sua matriz de confusão. Quanto mais próximo o método se encontrar do canto superior esquerdo do gráfico $(\text{TFP} = 0, \text{TVP} = 1)$, melhor ele pode ser considerado. Outros dois pontos importantes no espaço *ROC* são o $(\text{TFP} = 0, \text{TVP} = 0)$ e o $(\text{TFP} = 1, \text{TVP} = 1)$. O primeiro corresponde a um método que sempre apresenta um diagnóstico negativo enquanto que o segundo representa um método que sempre apresenta um diagnóstico positivo.

Para métodos *CAD* capazes de apresentar um valor contínuo entre 0 e 1, denominado de *score-P*, que representa um grau de certeza para previsão da diagnóstico positivo, é possível gerar um curva no

Tabela 2.4: Medidas derivadas da matriz de confusão binária.

Medida	Descrição
Acurácia $= \frac{VP+VN}{VP+VN+FP+FN}$	Proporção dos casos de teste diagnosticados corretamente.
Revocação, $r = \frac{VP}{VP+FN}$	Também é denominada de TVP (taxa de verdadeiros positivos) ou sensibilidade. Corresponde à fração de todos os casos positivos que foram corretamente diagnosticados pelo método *CAD* como positivos.
Especificidade $= \frac{VN}{VN+FP}$	Também é denominada de TVN (taxa de verdadeiros negativos). Corresponde a fração de todos os casos negativos que foram corretamente diagnosticados pelo método *CAD* como sendo negativos.
Taxa de FP $= \frac{FP}{FP+VN}$	Também é denominada de TFP (taxa de falsos positivos). Corresponde a fração de todos os casos negativos que foram incorretamente diagnosticados pelo método *CAD* como sendo positivos.
Precisão, $p = \frac{VP}{VP+FP}$	Fração dos casos que são de fato positivos dentre todos os casos diagnosticados pelo método *CAD* como positivos.
Medida-$F_1 = \frac{2}{1/r+1/p}$	Média harmônica das medidas de precisão e revocação.

espaço *ROC*. Para tanto, deve-se primeiramente ordenar os casos diagnosticados de acordo com seu valor de *score-P*. Define-se então um limiar de binarização que é utilizado para diagnosticar os casos como positivos e negativos a partir do *score-P*. Casos que apresentem um *score-P* menor e maior que o valor de limiar são diagnosticados, respectivamente, como negativos e positivos. Ao variar o valor do limiar de binarização entre 0 e 1, obtem-se uma curva no espaço *ROC*. A curva *ROC* pode ser utilizada, por exemplo, para se escolher um ponto de operação ideal do método de classificação ao se adotar o limiar de binarização que resulta no ponto mais próximo do canto superior direito do gráfico.

Adicionalmente, como ilustrado na figura 2.7, uma curva *ROC* pode ser utilizada para comparar a performance de diferentes métodos *CAD*. No exemplo, o método M1 é melhor que o método M2 quando a TFP é maior que 0,5 enquanto M2 é superior para TFP menores que 0,5. Desta maneira, a escolha de um método ou outro deve ser realizada levando em conta o custo de erros do tipo FP para a aplicação, não sendo possível afirmar de maneira genérica que um destes métodos é melhor que o outro. Note que a linha tracejada na diagonal ascendente do plano *ROC* corresponde a um método de comportamento estocástico.

2.7 Considerações Finais

Este capítulo descreveu como a mineração de imagens pode ser aplicada em bases de imagens médicas com o propósito de se desenvolver técnicas de diagnóstico auxiliado por computador. Foram apresenta-

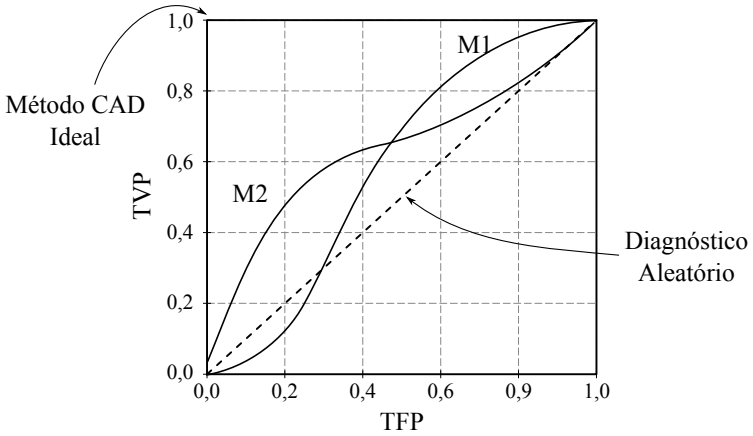


Figura 2.7: Curva *ROC* para dois métodos *CAD*.

dos, inicialmente, os principais conceitos relacionados à mineração de imagens, incluindo o algoritmo *Apriori* de mineração de regras de associação. Em seguida, foi apresentado o método *IDEA*, que sugere palavras chaves para compor um diagnóstico de uma imagem médica com base em regras de associação e que foi utilizado como base para o desenvolvimento deste projeto de mestrado. Por fim, foi realizada uma breve discussão sobre como realizar a avaliação de sistemas *CAD*, incluindo análise ROC. O próximo capítulo apresenta uma discussão sobre o processo de extração de características visuais de imagens, que corresponde a uma das principais etapas do processo de mineração de imagens.

Capítulo 3

Descritores de Forma

Enquanto que na mineração de dados tradicional os conjuntos de dados estão organizados em tabelas, nas quais as linhas são exemplos e as colunas são medidas de um determinado atributo, na mineração de imagens existe o desafio de se obter uma representação adequada da informação. Para este propósito, é necessário utilizar descritores para extrair e comparar as características visuais das imagens. Essas características são, em geral, representadas na forma de vetores de características que sumarizam atributos visuais e possibilitam medir a dissimilaridade entre duas imagens por meio de funções de distância. O processo de extração implica em redução de dimensionalidade, uma vez que a imagem, que pode ser vista como uma matriz bidimensional de níveis de cinza, apresenta um maior número de componentes que o vetor de características.

Formalmente, um vetor de características $\vec{v}_I = (v_1, v_2, \cdots, v_n)$ de uma imagem I pode ser definido como um ponto no espaço $\mathbb{R}^n$ onde n é o número de dimensões desse vetor [Torres 06]. A imagem em níveis de cinza I, por sua vez, é um par $(P_I, \vec{I})$ onde:

- P_I é um conjunto finito de pixels (pontos em $\mathbb{N}^2$ tal que $P_I \subset \mathbb{N}^2$), e
- $\vec{I}$: $P_I \rightarrow \mathbb{R}$ é uma função que associa cada pixel p em P_I a um escalar em $\mathbb{R}$ que representa a intensidade do nível de cinza.

Tomando esta definição, uma imagem pode ser representada por uma matriz de N linhas e M colunas na qual $I(x,y)$ denota o valor do pixel p na posição (x,y). Por fim, um descritor D de imagem é definido como um par $(\varepsilon_D, \delta_D)$ na qual:

- $\varepsilon_D : I \rightarrow \mathbb{R}^n$ é uma função que extrai o vetor de características $\vec{v}_I$ da imagem I. O espaço n-dimensional dos vetores de características $\vec{v}_I$ é denotado por $\mathcal{F}$ e denominado espaço de características.
- $\delta_D : \mathbb{R}^n \times \mathbb{R}^n$ é uma função de similaridade que computa a similaridade entre duas imagens como sendo o inverso da distância entre seus respectivos vetores de características.

A figura 3.1 ilustra o processo de extração e comparação das características visuais. Primeiramente, os vetores de características das imagens I_A e I_B são extraídos utilizando a função ε_D do descritor D. Um exemplo de vetor de características que poderia ser extraído nesta etapa é o histograma de níveis de cinza, que então seria entrada para função de similaridade δ_D. Uma possível escolha para função δ_D é a distância Euclidiana, que retorna um valor d de distância que corresponde à dissimilaridade entre I_A e I_B. Deve-se notar que é possível combinar descritores para compor um descritor mais complexo, capaz de descrever vários tipos de informações visuais presentes em uma imagem [Torres 05].

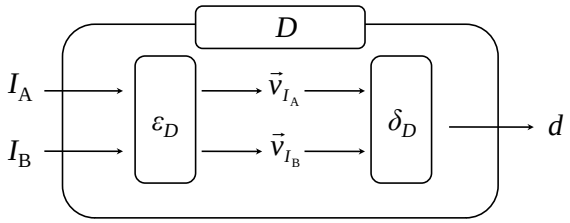


Figura 3.1: Diagrama ilustrando a extração e comparação das características das imagens I_A e I_B pelo descritor D.

As características visuais mais comuns empregadas na representação de imagens são aquelas definidas como primitivas [Liu 07], tais como cor, textura e forma. Descritores baseados em cor, em geral, fazem uso do histograma de níveis de cinza da imagem para obter sua representação [Kiranyaz 10]. Descritores de textura constituem uma medida do arranjo estrutural dos pixels em uma imagem. Algumas das técnicas mais conhecidas de extração de características de textura baseam-se nas wavelets [Wang 01], nos filtros de Gabor [Ma 99] e sumarizações das matrizes de co-ocorrência (conhecidas como descritores de Haralick) [Haralick 73]. Embora não exista uma definição formal bem estabelecida, as medidas de textura capturam essencialmente a granularidade e padrões repetitivos na distribuição dos pixels. Os descritores de forma, por terem sido de especial interesse para este trabalho, são discutidos em maiores detalhes na próxima seção.

3.1 O Processo de Extração de Descritores de Forma

Ao observar uma imagem de radiografia, um especialista tende a descrevê-la com informações de valor semântico, tais como a ocorrência de ossos fraturados ou tumores. Na mineração de imagens, em contrapartida, faz-se uso de informações de baixo nível (sintáticas) tais como a distribuição de cor ou níveis de cinza para caracterizar uma imagem. Esta diferença no modo como um usuário interpreta as informações presentes em uma imagem e os atributos visuais que um descritor pode extrair da mesma resulta em um problema conhecido como "lacuna semântica" (*semantic gap*) [Zhuang 07, Fischer 08].

Para amenizar este problema, o projetista de um sistema de mineração de imagens deve recorrer a características de baixo nível que possuam o máximo de correlação possível com as informações semânticas da imagem [Kinoshita 07]. Um dos atributos visuais que melhor se aproxima à percepção humana é a característica de forma. Seres humanos são capazes de reconhecer objetos apenas por seu contorno, como se pode perceber pelas silhuetas mostradas na figura 3.2. Apesar de não ser fornecida nenhuma informação sobre cor, contexto ou textura, os objetos podem ser facilmente identificados [Costa 01].

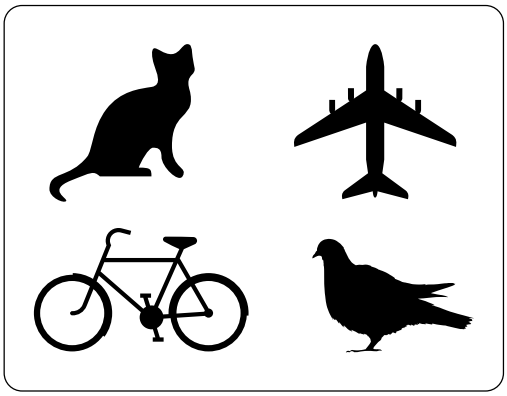


Figura 3.2: Objetos que podem ser reconhecidos somente por sua forma (fonte: *Wikimedia Commons*, `http://commons.wikimedia.org`).

Para aplicar a função de extração de características ε_D em formas, é necessário que as imagens sejam previamente segmentadas. A segmentação tem como objetivo identificar regiões dentro da imagem que compartilhem determinadas características em comum, tais como textura, cor ou níveis de cinza. Trata-se de uma etapa fundamental, uma vez que é nela que os objetos a serem caracterizados por descritores de forma são evidenciados. Em [Xu 00, Zhang 08], o problema da segmentação é formalizado definindo-se P_I como sendo o conjunto de todos os pixels pertencentes a uma imagem I e $S_k \subset P_I$ como sendo os conjuntos de pixels que formam as regiões na imagem segmentada que devem satisfazer a relação:

$$P_I = \bigcup_{k=1}^{N} S_k \tag{3.1}$$

onde $S_i \cap S_j = \emptyset$ para $i \neq j$ e N é o número de regiões obtidas. Uma vez que os descritores a serem apresentados trabalham com imagens binárias, o processo de segmentação gera duas regiões S_1 e S_2 que podem ser representadas por uma imagem binária na qual cada pixel p em P_I pode assumir os valores 1 e 0, ou seja, fundo e objeto respectivamente.

A tabela 3.1 descreve resumidamente algumas das abordagens de segmentação existentes na literatura. Deve-se notar que a quantidade de métodos de segmentação é bastante vasta. No entanto, por não terem sido o foco de pesquisa deste projeto de Mestrado, a etapa de segmentação de imagens não será discutida em detalhes.

É importante observar que a representação de um objeto por meio de uma imagem binarizada, resultante do processo de segmentação, não é adequada para alguns dos métodos de extração de características de forma. Por este motivo, alguns descritores de forma incorporam uma etapa de obtenção de representações intermediárias de forma que é discutida em maiores detalhes na seção 3.2.

A figura 3.3 ilustra as etapas que compõem o processo de extração de características de forma. Inicialmente é realizada uma segmentação da imagem que é então sucedida pela obtenção de uma representação intermediária de forma. No exemplo, fez-se uso do contorno paramétrico. Por fim, é realizada a extração do vetor de características que é utilizado no processo de mineração de imagens.

Tabela 3.1: Abordagens de segmentação de imagem.

Abordagem de Segmentação	**Descrição**	**Principais Características**
Limiarização	Pixels com níveis de cinza que pertençam a uma mesma faixa de valores são agrupados em uma mesma região. As faixas de valores são delimitadas por limiares que podem ser definidos de maneira supervisionada ou automática. O método de Otsu [Otsu 79] encontra de maneira automática o valor de limiar que torna a distribuição dos níveis em cada região o mais homogênea possível.	• Baixo custo computacional. • Apresenta bons resultados quando cada região da imagem apresenta diferentes intervalos de níveis de cinza.
Detecção de Bordas	Consiste em encontrar as regiões da imagem onde a variação de níveis de cinza ocorre de maneira abrupta. Métodos clássicos de detecção de bordas incluem o filtro de Marr-Hildreth [Marr 80] e o detector de Canny [Canny 86], ambos baseados em operadores de derivação e em algoritmos de enlace. Modelos de contornos ativos (ou *snakes*) [Kass 88] são curvas suaves (*splines*) que se ajustam a bordas e linhas de um objeto com base em forças definidas a partir da "topologia" da imagem.	• A presença de ruídos na imagem pode causar falsas bordas, tornando o processo de aplicação de algoritmos de enlace mais complexo. • A posição e forma iniciais das *snakes* devem ser bem próximas da borda do objeto ao qual se deseja que a curva se adapte.
Transformada *Watershed*	Na transformada *watershed* [Beucher 79] uma imagem em níveis de cinza é observada como um relevo. Considerando o comportamento de uma gota de água que caia sobre esse relevo, existirão pontos para os quais será possível determinar um mínimo local para onde a gota irá drenar. No entanto, haverá pontos para os quais a gota de água poderia drenar para dois diferentes pontos de mínimos. Este último conjunto de pontos define linhas de divisão que são utilizadas para segmentar a imagem.	• Ruídos ou estruturas complexas podem causar super-segmentação da imagem. • Marcadores, que definem artificialmente mínimos locais na imagem, podem ser utilizados para controlar o número de regiões e evitar a super-segmentação.
EM/MPM	O EM/MPM [Comer 94] é um algoritmo iterativo que utiliza um campo aleatório de Markov para modelar os rótulos de região dos pixels. De maneira alternada, a técnica MPM (*Maximization of the Posterior Marginals*) é utilizada para classificar os pixels e o algoritmo EM, de maximização da esperança, é utilizado para estimar os parâmetros da função probabilidade dos rótulos dos pixels.	• Não supervisionado. Apenas exige que seja determinada a quantidade de regiões resultante do processo de segmentação. • Segmentação por textura.

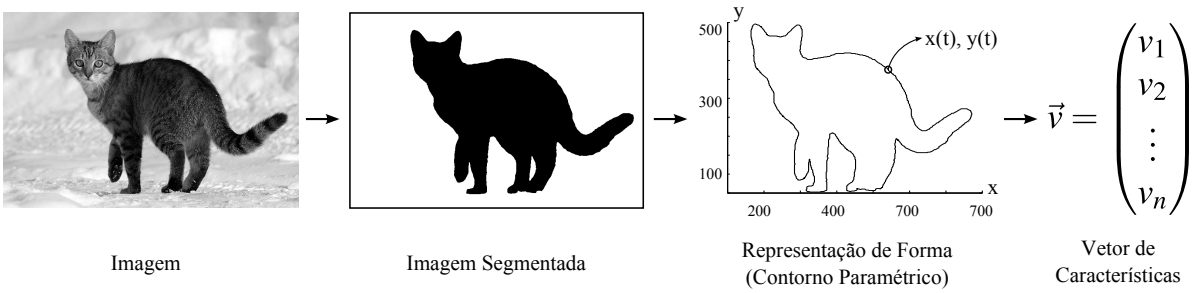


Figura 3.3: Etapas do processo de extração de características de forma.

3.2 Representação de Formas

As representações de formas podem ser divididas em três categorias [Costa 01]. As duas primeiras são baseadas em contorno e em região, que serão discutidas ao longo desta seção. A terceira categoria refere-se às representações baseadas em transformadas que serão abordadas na seção 3.3 por também poderem ser utilizadas diretamente como descritores de forma. De fato, uma série de representações intermediárias de forma podem ser utilizadas diretamente como descritores. Um exemplo adicional é o código de cadeia (*chain code*), que pode tanto ser utilizado como uma representação intermediária quanto como um descritor de forma.

3.2.1 Representações Baseadas em Contorno

Representações baseadas em contorno exploram as informações contidas na borda de um objeto. Uma maneira rudimentar de representar uma forma através desta abordagem consiste na obtenção do conjunto de pontos pertencentes à sua borda. Uma vez que nesta representação os pontos não apresentam uma relação de ordem entre si, sua aplicação no contexto de extração de descritores de forma é bastante limitada sendo possível, no entanto, extrair medidas tais como o centróide e o eixo principal da forma.

Uma representação de contorno mais poderosa pode ser obtida definindo uma relação de ordem entre os pontos que compõem a fronteira da forma. Para tanto, é escolhido um ponto inicial arbitrário ao longo da borda de um objeto que é então percorrida no sentido horário ou anti-horário. O resultado deste processo é uma lista de pontos denominada de contorno paramétrico que é análoga à representação paramétrica de uma curva na geometria diferencial. A figura 3.4(a) ilustra o processo de obtenção da representação resultando no seguinte conjunto de pontos ordenados:

$$(6,3),(5,3),(4,3),(4,4),(4,5),(4,6),(3,6),(2,6),(2,5),(2,4),(2,3),(2,2),$$
$$(2,1),(3,1),(4,1),(5,1),(6,1),(6,2)$$

É importante notar que a mudança no sentido (horário e anti-horário) em que se percorre a borda do objeto resulta em uma série de implicações no processo de extração de características. Por exemplo, a curvatura tem seu sinal invertido ao se escolher uma orientação diferente. Por este motivo, é importante adotar uma convenção para a orientação do contorno e utilizá-la de maneira consistente. Adicionalmente, o contorno paramétrico pode ser visto como um par de sinais discretos $x(n)$ e $y(n)$, que representam as coordenadas x e y dos pontos que compõem o contorno conforme ilustrado na figura 3.4(b).

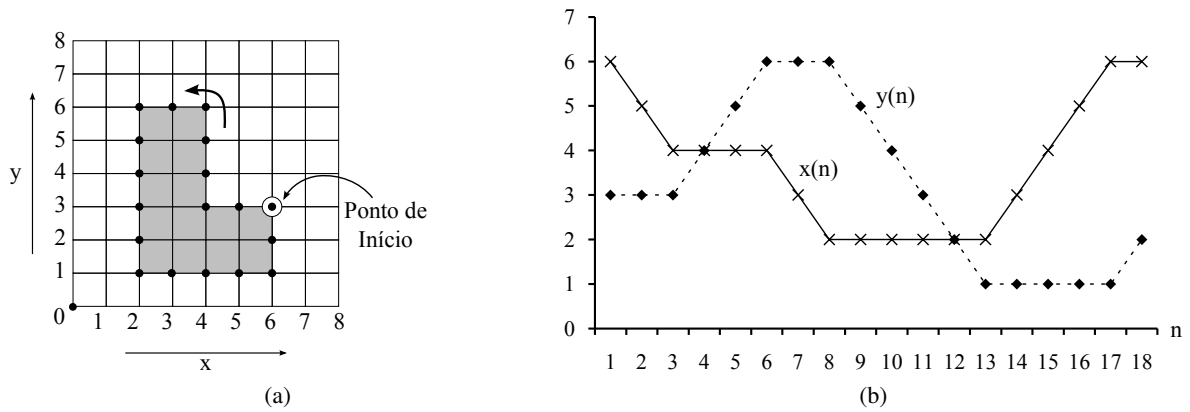


Figura 3.4: Obtenção da representação de forma por contorno paramétrico. Figura adaptada de [Costa 01]. (a) O contorno paramétrico é obtido seguindo o contorno no sentido anti-horário partindo do ponto $(6,3)$. (b) Sinais discretos definidos pela representação paramétrica da forma.

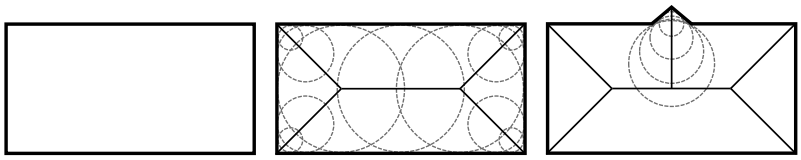
Uma abordagem alternativa à representação por curvas paramétricas consiste em aproximar a borda de um objeto por um conjunto de primitivas geométricas. Por exemplo, pode-se segmentar o contorno e aproximar os segmentos de curva por segmentos de reta, resultando em uma representação poligonal da forma. Os pontos que definem a segmentação do contorno são denominados pontos dominantes. Um método para se encontrar tais pontos consiste em determinar os pontos da curva que possuem um maior valor para o módulo da medida de curvatura. Alternativamente, pode-se definir uma função de erro, como por exemplo, a maior distância entre o contorno e as arestas do polígono, e encontrar o conjunto de pontos dominantes que a minimize.

3.2.2 Representações Baseadas em Região

Enquanto os métodos de representação de forma baseados em contorno fazem uso somente da informação presente na fronteira da forma de um objeto, métodos baseados em regiões fazem uso de toda a informação contida na região interna de um objeto. Desta maneira, a própria imagem binarizada resultante do processo de segmentação, na qual os pixels de valor 1 correspondem ao fundo e os pixels de valor 0 correspondem à região do objeto, pode ser considerada uma representação de forma baseada em região. Um exemplo de representação baseada em região é o esqueleto multiescala. Informalmente, um esqueleto pode ser definido como uma representação que se encontra na região central das várias partes que compõem uma forma e é tão estreita quanto possível. De fato, a idéia básica de um esqueleto consiste em eliminar informações redundantes, conservando somente a estrutura topológica do objeto que pode ser utilizada na tarefa de extração de características de forma [Zhang 04].

Um dos possíveis métodos que podem ser utilizados para obtenção do esqueleto de forma consiste em utilizar a Transformada do Eixo Médio *MAT* (*Medial Axis Transform*) [Blum 67]. A *MAT* de uma forma corresponde às posições ocupadas pelos centros das circunferências que: (i) são bitangentes, ou seja, tocam a curva em dois pontos e (ii) são completamente internas à forma. A grande limitação da *MAT* é sua susceptibilidade a ruídos. Por exemplo, uma pequena saliência no contorno pode resultar em um novo ramo no esqueleto. A figura 3.5 ilustra uma forma simples e seu respectivo esqueleto resultante da aplicação da *MAT*. As circunferências tracejadas em 3.5b correspondem aos círculos bitangentes internos ao contorno cujos centros definem a região do esqueleto. A figura 3.5c mostra como uma

pequena saliência na borda pode resultar em um novo ramo no esqueleto.



(a) (b) (c)

Figura 3.5: Obtenção de esqueletos através da Transformada do Eixo Médio (*MAT*) [Costa 01]. (a) Uma forma simples. (b) Seu respectivo esqueleto e as circunferências bitangentes internas. (c) Pequenas saliências na borda da forma podem causar o surgimento de novos ramos no esqueleto.

Para contornar o efeito do ruído da *MAT*, é proposto em [Costa 99] um método para obtenção de esqueletos multiescala. Inicialmente são atribuídos rótulos à borda da forma. Esses rótulos são propagados através de dilatações exatas e utilizados para gerar uma imagem de diferenças onde cada pixel corresponde à máxima diferença entre seu rótulo e de seus quatro vizinhos. Os esqueletos multiescala podem ser obtidos simplesmente realizando uma limiarização da imagem de diferenças. No entanto, conforme o próprio autor de [Costa 99] menciona em [Falcão 02], o custo computacional para se propagar os rótulos utilizando o método de dilatações exatas é excessivo mesmo para imagens de tamanho médio. Em [Falcão 02] é proposta uma abordagem baseada na Transformada Imagem-Floresta (*Image Foresting Transform - IFT*) para tornar o processo de propagação de rótulos mais eficiente.

Outras abordagens para extração de representações de forma baseada em regiões incluem métodos que fazem uso de matrizes de forma (*shape matrix*), *quad-trees*, dendrogramas e retângulos mínimos delimitadores [Zhang 04, Costa 01].

3.3 Revisão Bibliográfica de Descritores de Forma

Uma vez que a quantidade de descritores de forma existente na literatura é bastante vasta, não seria possível nesta dissertação realizar um levantamento completo de todos eles, por não ser esse o principal tema da pesquisa realizada. Por este motivo, optou-se por apresentar os métodos de extração que serviram como base para o desenvolvimento deste projeto de Mestrado.

Primeiramente, na seção 3.3.1, são apresentados os descritores de forma básicos que são relacionados à aspectos geométricos da forma de um objeto. Esta classe de descritores apresenta um baixo custo computacional de extração e pode ser utilizada para diferenciar classes de objetos que apresentem formas significativamente diferentes. No entanto, se as diferenças entre as formas forem sutis, o poder de discriminação dos descritores básicos é insuficiente. No contexto de imagens médicas, por exemplo, massas tumorais são, em geral, maiores que microcalcificações e, por este motivo, descritores básicos de forma poderiam ser aplicados para separar essas duas classes de formas.

Na seção 3.3.2 são apresentados os descritores de forma baseados em transformadas, mais especificamente os descritores de Fourier. Os descritores de Fourier apresentam uma série de importantes características, sendo estas (i) processo de extração eficiente; (ii) possibilidade de realizar normalização quanto a rotação, escala e translação do objeto e (iii) possibilidade de capturar tanto características globais quanto locais da forma [Zhang 04].

Por fim, na seção 3.3.3, são apresentados os descritores de forma baseados na dimensão fractal. Fractais são estruturas que apresentam a propriedade de auto-similaridade. Muitas estruturas naturais que são alvo da análise de sistemas *CAD* também apresentam esta propriedade (ainda que em escalas limitadas) e, por este motivo, podem ser modeladas como fractais. Desta maneira, a dimensão fractal pode ser utilizada como uma medida de complexidade de tais estruturas. De fato, o método de extração *FFS* (*Fast Fractal Stack*) desenvolvido neste trabalho e descrito no capítulo 4 emprega a análise fractal para descrever a forma de imagens médicas.

3.3.1 Descritores Básicos

Uma das maneiras mais simples de se caracterizar um objeto no contexto de mineração de imagens consiste na utilização de medidas relacionadas a aspectos métricos de sua forma, como por exemplo seu tamanho. Conforme mencionado em [Zhang 04], essa classe de descritores apresenta um baixo poder de discriminação. No entanto, mesmo medidas simples de forma podem ser úteis em situações específicas onde há uma grande variação nas formas dos objetos.

O total de pixels internos a um objeto na representação binária resultante do processo de segmentação pode ser utilizado para calcular sua área **A**. Uma vez que esse valor é dado em número de pixels, pode ser necessário em determinadas aplicações realizar uma conversão da medida para as unidades originais da cena, como, por exemplo, em metros ou centímetros. Para tanto, é importante considerar a resolução do processo de amostragem. Uma outra medida que pode ser extraída da imagem binária é o centróide, que consiste na média das coordenadas de todos os pixels internos ao objeto.

A representação por contorno paramétrico também pode ser utilizada para obtenção de descritores básicos de forma. Sendo $x(n)$ e $y(n)$ os sinais discretos definidos pela representação paramétrica do contorno (ver figura 3.4), o perímetro **P** é dado pela equação:

$$\mathbf{P} = \sum_{n=1}^{N-1} \sqrt{(x(n) - x(n+1))^2 + (y(n) - y(n+1))^2} \tag{3.2}$$

onde N corresponde ao número de elementos nos sinais $x(n)$ e $y(n)$.

O contorno paramétrico pode ser utilizado ainda para se calcular o diâmetro, que é definido em [Costa 01] como a maior distância entre dois pontos do contorno paramétrico, e também a distância média, mínima e máxima do contorno ao centróide.

Uma vez que em geral a complexidade de uma forma apresenta uma relação com sua capacidade de cobrir o espaço [Costa 01], os descritores básicos de forma podem ainda ser empregados para se derivar medidas relacionadas à complexidade da forma de um objeto. Uma dessas medidas é a circularidade, que corresponde à razão entre o quadrado do perímetro e a área. Outra medida é o alongamento, dado pela razão entre o comprimento dos eixos principais de uma forma. Os eixos principais são um par de eixos normais entre si formados pelo eixo maior (direção na qual os pontos que compõem um objeto estão mais dispersos) e o eixo menor. A figura 3.6 mostra uma forma e seus respectivos eixos principais. Uma técnica que pode ser utilizada para obter os eixos principais consiste em tratar as coordenadas x e y de cada ponto interno ao objeto como variáveis aleatórias X e Y e calcular sua matriz de covariância Σ dada por:

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Cov}(Y,Y) \end{bmatrix} \tag{3.3}$$

onde $\mathrm{Cov}(A,B)$ é a covariância das variáveis aleatórias A e B. Os autovetores de Σ correspondem aos eixos principais da forma e o alongamento pode ser calculado como a razão entre o maior e menor o autovalores de Σ.

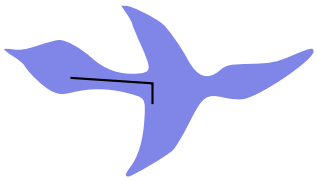


Figura 3.6: Forma e seus eixos principais, que podem ser utilizados para calcular o descritor de alongamento [Costa 01].

3.3.2 Descritores de Fourier

A transformada de Fourier é uma operação matemática que decompõe um sinal nas freqüências que o constituem, ou seja, dada uma função, sua transformada de Fourier consiste em uma combinação linear de senos e cossenos ou exponenciais complexas. Por exemplo, um acorde produzido por um instrumento musical pode ser visto como um sinal no tempo e sua transformada de Fourier corresponde às notas que o compõem. Uma série de descritores de forma, denominados de descritores de Fourier, podem ser derivados da transformada discreta de Fourier (DFT) do contorno de uma forma.

Em [Granlund 72] os sinais discretos $x(n)$ e $y(n)$ definidos pela representação paramétrica da forma (figura 3.4b) são utilizados para compor um sinal complexo $z(n) = x(n) + jy(n)$. A expansão de $z(n)$ em uma série de Fourier é dada por:

$$Z(k) = \frac{1}{N}\sum_{n=0}^{N-1} z(n)\exp\left(-j\frac{2\pi}{N}nk\right), \quad k = -\frac{N}{2},\cdots,-1,0,1,\cdots,\frac{N}{2}-1 \tag{3.4}$$

onde N corresponde ao número de pontos na representação paramétrica do contorno. Os coeficientes $Z(k)$ são conhecidos como descritores de Fourier e podem ser obtidos de maneira eficiente pelo algoritmo FFT (*Fast Fourier Transform*). É interessante notar que o coeficiente $Z(0)$ corresponde ao centróide da forma.

Uma característica importante dos descritores de Fourier é sua capacidade de caracterizar o quanto uma forma é suave ou irregular. As figuras 3.7(a) e 3.7(c) mostram os contornos de uma massa tumoral maligna e benigna e as respectivas magnitudes dos descritores de Fourier em escala logarítimica (figuras 3.7(b) e 3.7(d)). Pode-se perceber que para a massa maligna, que apresenta um contorno irregular, ocorre uma maior presença de energia nas freqüências altas quando se compara a magnitude de seus descritores de Fourier com a magnitude dos descritores de Fourier da massa benigna, que possui um contorno mais suave.

Em [Shen 94] é apresentada a medida de fator de forma, denotada por **ff**, que é derivada dos descritores de Fourier. A **ff** é invariante a translação, rotação e escala e seu valor aumenta conforme a borda do objeto se torna mais irregular. O método para sua obtenção consiste, inicialmente, em normalizar

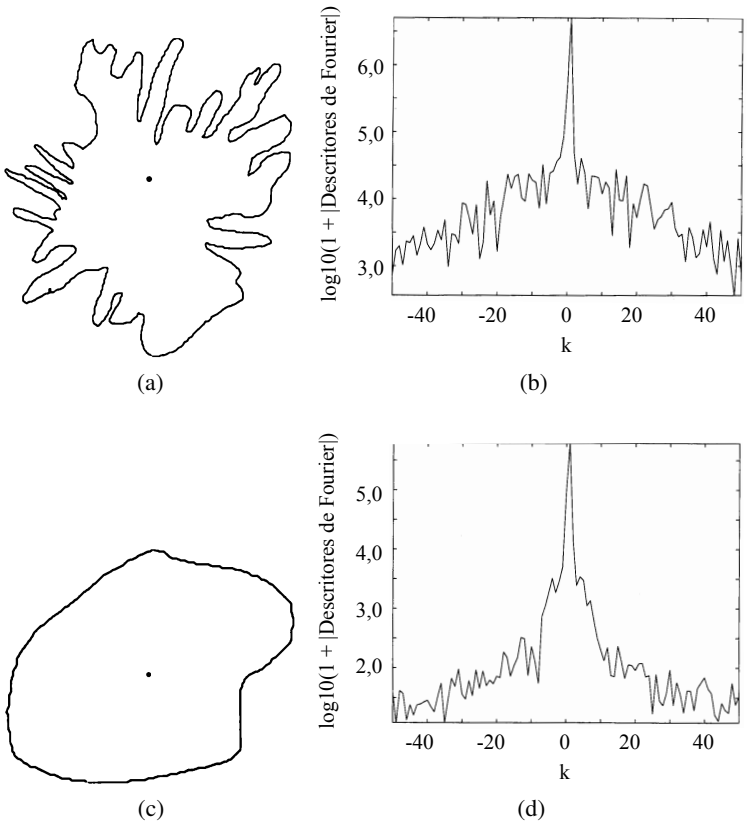


Figura 3.7: Extração de descritores de Fourier de contornos de massas tumorais [Rangayyan 04]. (a) e (c) mostram os contornos de uma massa tumoral maligna e benigna e as respectivas magnitudes dos descritores de Fourier em escala logarítimica, (b) e (d).

os descritores de Fourier. A normalização quanto a translação é obtida atribuindo o valor zero para o coeficiente $Z(0)$, que corresponde à posição do centróide. Para tornar os descritores invariantes quanto à escala, cada coeficiente é dividido pela magnitude de $Z(1)$. Os descritores normalizados de Fourier $Z_o(k)$ são definidos como:

$$Z_o(k) = \begin{cases} 0, & k = 0; \\ \frac{Z(k)}{|Z(1)|}, & \text{caso contrário.} \end{cases} \tag{3.5}$$

A partir dos coeficientes normalizados $Z_o(k)$, a medida de fator de forma **ff** é calculada como:

$$\mathbf{ff} = 1 - \frac{\sum_{k=-N/2+1}^{N/2} (|Z_o(k)|/|k|)}{\sum_{k=-N/2+1}^{N/2} |Z_o(k)|} \tag{3.6}$$

Como se pode perceber pela análise da equação, cada coeficiente $Z_o(k)$ no numerador da fração é ponderado pelo fator $1/|k|$. Ou seja, quanto maior a freqüência associada a um coeficiente, menor será seu peso no cálculo de **ff**. Uma vez que os ruídos do contorno de uma forma tendem a se concentrar nas freqüências mais altas, a ponderação utilizada torna a **ff** menos sensível a ruídos.

Assim como grande parte dos descritores de forma, a extração de descritores de Fourier necessita que o objeto para o qual a forma será analisada seja evidenciado por um processo de segmentação (figura 3.3). Em [Timm 10] é proposto um conjunto de descritores estatísticos de Fourier, denominados SFDs

(*Statistical Fourier Descriptors*). A grande vantagem deste método é que ele dispensa a segmentação da imagem em objetos conexos para realizar o processo de extração de características de forma.

A figura 3.8 ilustra o processo de extração dos SFDs. Inicialmente, a imagem de entrada é decomposta em uma pilha de imagens binárias através de um processo de decomposição por limiarização proposto em [Chen 95]. Esse processo corresponde à parte **(A)** da figura 3.8 e é realizado através de consecutivas segmentações por limiarização para diferentes valores de limiar. Em seguida, na etapa **(B)**, os componentes conexos brancos c^w e pretos c^b são extraídos. Para cada componente c, é computado o conjunto de descritores de Fourier local, denotado por g_c^*. A etapa **(D)** tem como saída dois conjuntos de descritores por imagem binarizada, denotados por $\bar{g}_c^w$ e $\bar{g}_c^b$ que correspondem, respectivamente, a combinação dos descritores locais para os componentes brancos e pretos. Essa combinação é realizada por meio do cálculo de medidas estatísticas dos descritores locais obtidos na etapa **(C)**. Por fim, na etapa **(E)**, um processo similar é utilizado para combinar os descritores $\bar{g}_c^w$ e $\bar{g}_c^b$ de cada imagem binarizada em um único vetor SFDs, denotado pela letra **h**.

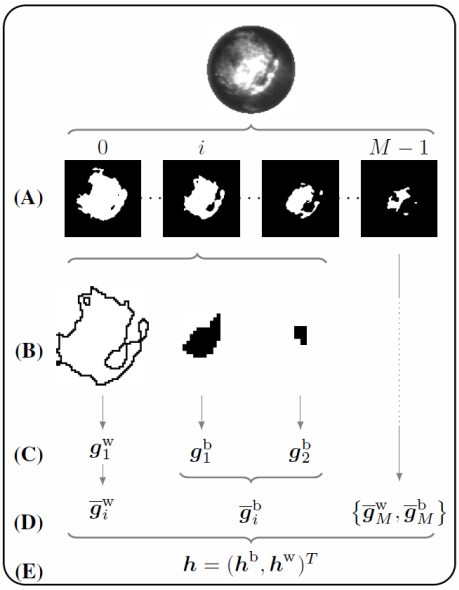


Figura 3.8: Processo de extração dos descritores estatísticos de Fourier [Timm 10].

3.3.3 Descritores Baseados em Análise Fractal

Fractais são estruturas naturais ou artificiais que apresentam a propriedade de auto-similaridade, ou seja, objetos compostos por muitas partes que são, cada uma, similares ao fractal como um todo [Mandelbrot 83]. Ao se tomar um triângulo equilátero e remover o triângulo interno composto pela união de seus três pontos médios, tem-se como resultado três novos triângulos equiláteros cujos lados têm comprimento igual à metade do original. Removendo o triângulo interno de cada um dos três triângulos restantes e repetindo esse processo indefinidamente, como ilustrado na figura 3.9, têm-se como resultado o triângulo de Sierpinski, uma figura geométrica fractal, que apresenta perímetro infinitamente grande e área nula.

De fato, os fractais podem apresentar características bastante incomuns. Por exemplo, objetos da geometria Euclidiana apresentam valores naturais para sua dimensão. Ou seja, um ponto possui dimensão

Figura 3.9: As cinco primeiras iterações do processo construção do triângulo de Sierpinksi.

zero, uma reta possui dimensão um e um plano, dimensão dois. Fractais, no entanto, podem possuir dimensões com valores fracionários. O triângulo de Sierpinski, por exemplo, possui uma dimensionalidade com valor entre um e dois.

Para fractais exatamente auto-similares, gerados por processos iterativos infinitos e bem definidos, existe uma fórmula simples para o cálculo da dimensão fractal, denotada por $\mathcal{D}$. Segundo [Schroeder 92], dado um fractal cuja regra de criação gere M réplicas do fractal original de forma que cada réplica seja uma versão do mesmo em escala $1:f$, sua dimensão fractal $\mathcal{D}$ é dada por:

$$\mathcal{D} = \frac{\log M}{\log f} \tag{3.7}$$

Para o triângulo de Sierpinski, por exemplo, a regra de construção gera três réplicas do triângulo original em uma escala $1:2$, ou seja $M = 3$ e $f = 2$. Com isso, tem-se que $\mathcal{D} = \log 3/\log 2 \approx 1,58$.

No contexto de análise de formas, a dimensão fractal pode ser utilizada para quantificar a complexidade de objetos que apresentem estruturas auto-similares em diferentes escalas [Costa 01]. É importante notar, no entanto, que o cálculo da dimensão fractal dada pela equação 3.7 não pode ser aplicado a fractais que não possuem uma regra de construção bem definida, como é o caso das formas presentes em imagens reais. Outra maneira de se calcular a dimensão fractal é através da dimensão de Hausdorff [Schroeder 92], denotada por $\mathcal{D}_0$. Para tanto, um objeto que possua uma dimensão Euclidiana E é dito ser preenchido por $N(\gamma)$ cubos de dimensão E e lados de comprimento γ. O valor de $\mathcal{D}_0$ pode ser obtido pela expressão:

$$\mathcal{D}_0 = \lim_{\gamma \to 0} \frac{\log N(\gamma)}{\log \gamma^{-1}} \tag{3.8}$$

Para um objeto representado por uma imagem binária, um valor aproximado para $\mathcal{D}_0$ pode ser obtido através do algoritmo de contagem de caixas. Inicialmente, a imagem é dividida em uma grade composta por quadrados de tamanho $\gamma \times \gamma$ e então é feita a contagem $N(\gamma)$ dos quadrados que contém ao menos uma parte da forma. Ao variar o valor de γ é possível criar uma curva $\log N(\gamma) \times \log \gamma^{-1}$ que pode ser aproximada por uma reta. O valor do coeficiente de inclinação desta reta corresponde ao valor aproximado de $\mathcal{D}_0$. Na figura 3.10, adaptada de [Torres 04], é mostrada uma aproximação por uma imagem binária (figura 3.10a do fractal conhecido como estrela de Koch. A curva gerada pelo método de contagem de caixas e sua aproximação por uma reta são mostrados na figura 3.10b.

É mostrado em [Rangayyan 07] que tumores de mama podem ser caracterizados pela dimensão fractal de suas bordas. Uma vez que tumores malignos possuem padrões complexos e irregulares, a dimensão fractal de sua borda tende a apresentar valores mais altos que a dimensão fractal das bordas de tumores benignos. O cálculo da dimensão fractal foi realizado utilizando tanto a representação em duas dimensões da borda quanto o sinal em uma dimensão de distância do ponto da borda ao centróide. A

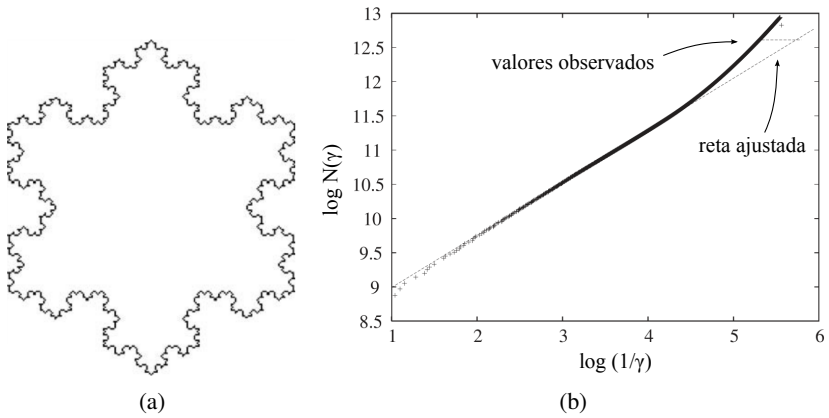


Figura 3.10: Método da contagem de caixas [Torres 04]. (a) Aproximação da estrela de Koch por uma imagem binária. (b) Curva $\log N(\gamma) \times \log \gamma^{-1}$ e sua aproximação por uma reta. O valor do coeficiente de inclinação da reta corresponde à uma aproximação da dimensão fractal de Hausdorff.

avaliação dos descritores foi realizada utilizando curvas ROC com uma base de 111 contornos de massas tumorais. Foi observado que o descritor de dimensão fractal pode ser utilizado para complementar outros descritores de forma.

Outro descritor de forma baseado na geometria fractal utilizado em trabalhos recentes [Bruno 08, Backes 10] é a dimensão fractal multiescala, que pode ser utilizada para medir a correlação existente entre a interface de um objeto e o espaço ou área que ele ocupa. Em [Torres 04] é proposto um método que analisa a correlação interface/área de um objeto por meio de dilatações exatas apoiada pela IFT [Falcão 02]. Dado um conjunto S de pontos representados por suas coordenadas (x,y), sua dilatação exata por um raio r, denotada como S_r, é definida como sendo a união de todos os discos de raio r centrados em cada um dos pontos em S. Supondo que S corresponda ao conjunto $\mathcal{C}$ de pontos que compõem a borda de um objeto, conforme se aumenta o valor de r têm-se como resultado uma versão cada vez mais simplificada do objeto.

Uma maneira eficiente de se obter a dilatação exata de uma forma por um raio r consiste em calcular a transformada de distância da imagem por meio do algoritmo proposto em [Falcão 02] e então realizar um processo de limiarização do mapa de distâncias resultante utilizando r como valor de limiar. Sendo $A(r)$ a área de S_r, variando o valor de r é possível gerar uma curva $\log A(r) \times \log r$, denominada de função de área logarítmica. O coeficiente angular da reta obtida por interpolação linear da curva corresponde a uma aproximação da dimensão fractal de Minkowski-Bouligand e os descritores de dimensão fractal multiescala, por sua vez, são amostras da derivada da função de área logarítmica.

3.4 Considerações Finais

Neste capítulo foram apresentados os conceitos fundamentais relativos à descritores de imagens e extração de características, com o objetivo de fornecer uma visão geral sobre a metodologia para a representação de imagens. Existem diversas maneiras de descrever formas, sendo encontrada uma vasta gama de métodos para esta tarefa na literatura. Neste capítulo, foram discutidos alguns dos métodos mais relevantes para o projeto, como os descritores de Fourier e os descritores baseados em Dimensão Fractal.

Parte II

Trabalhos Desenvolvidos

Capítulo 4

Extração de Características pelo Método *FFS*

Neste capítulo é descrito um novo método de extração de características desenvolvido durante este projeto de Mestrado: o *Fast Fractal Stack*, ou *FFS*. O algoritmo de extração consiste em decompor uma imagem em níveis de cinza em uma pilha de imagens binárias a partir das quais valores de dimensão fractal das imagens binarizadas são computados, resultando em um vetor de características compacto e altamente descritivo. Os resultados obtidos com o *FFS* para a tarefa de classificação de doenças pulmonares difusas em imagens de tomografias (*CT*) foram publicados no *ACM Workshop on Medical Multimedia Analysis and Retrieval* (*MMAR 2011*) que ocorreu juntamente com a conferência *ACM Multimedia 2011*. A abordagem proposta apresentou um desempenho superior quando comparada com outros algoritmos de extração de características. Adicionalmente, o algoritmo de extração do *FFS* é eficiente, com um custo computacional linear com respeito ao tamanho da imagem de entrada. O restante deste capítulo está organizado da seguinte maneira. A motivação para o método desenvolvido é apresentada na seção 4.1. A seção 4.2 descreve o método de extração do *FFS*. Resultados de experimentos são interpretados e discutidos na seção 4.3. As conclusões finais sobre o *FFS* são apresentadas na seção 4.4.

4.1 *Fast Fractal Stack - FFS*

O diagnóstico auxiliado por computador (*CAD*) de doenças pulmonares difusas (DPD) é um tópico de grande importância no campo de imagens de tomografia computadorizada de alta resolução (TCAR) [Uchiyama 03, Silva 09, Huber 10, Sluimer 06]. Essa importância pode ser atribuída ao rápido progresso nas tecnologias de aquisição de imagens de tomografia e também ao fato de que a interpretação de imagens de tomografia do pulmão de pacientes afetados com DPDs, mesmo para um radiologista experiente, é uma tarefa desafiadora e trabalhosa [Depeursinge 10]. Por este motivo, sistemas *CAD* confiáveis poderiam reduzir o trabalho manual dos radiologistas e evitar biópsias pulmonares desnecessárias.

Conforme discutido no capítulo 2, um sistema *CAD* típico consiste na extração de características

visuais relevantes de imagens na forma de vetores de características que são utilizados como entrada para classificadores. Devido ao problema conhecido como "lacuna semântica" (seção 3.1), que corresponde à diferença entre a percepção da imagem pelo especialista médico e suas características automaticamente extraídas, um aspecto desafiador da tarefa de extração de características é a obtenção de um conjunto de características que seja capaz de representar de maneira sucinta e eficiente o conteúdo visual de imagens médicas, apoiando o especialista médico no processo de tomada de decisão. Uma maneira de alcançar este objetivo seria extrair o maior número possível de características da imagem para melhor descrever seu conteúdo visual. No entanto, usar um grande número de características resultaria em um problema conhecido como maldição da dimensionalidade [Kriegel 09], no qual a significância e informações de cada característica diminui, fazendo com que o processo de classificação seja impreciso e custoso computacionalmente.

Desta maneira, é importante identificar e remover características redundantes ou irrelevantes. Seleção de atributos e transformação de características são duas técnicas que podem ser empregadas para este propósito. Na seleção de atributos o algoritmo de classificação (ou uma métrica baseada nas características do conjunto de dados) é utilizado para avaliar e selecionar um subconjunto de características a partir do conjunto original. Um exemplo de método de seleção de características é o *CFS* (*Correlation Feature Selection*)[Hall 00]. O *CFS* utiliza a performance preditiva das características e suas intercorrelações para encontrar um subconjunto de características que melhore o poder preditivo do classificador. Já os métodos de transformação, como o *PCA* (*Principal Component Analysis*), são capazes de gerar novas características que podem ser ordenadas por seu poder descritivo. Deste modo, é possível reduzir a dimensionalidade do conjunto de dados (número de atributos) descartando aqueles que são menos descritivos.

A desvantagem da seleção de atributos e transformação de características está no fato que demandam um custo computacional extra. Como uma abordagem alternativa para tratar de ambos os problemas, ou seja, da lacuna semântica e da maldição da dimensionalidade, foi desenvolvido neste projeto de Mestrado um novo método baseado na geometria fractal denominado *Fast Fractal Stack* (*FFS*) que extrai um vetor de características compacto e altamente descritivo de imagens em níveis de cinza. O *FFS* consiste em duas etapas principais:

1. Aplicação de uma técnica de particionamento de imagens (decomposição em pilha de imagens binárias, descrita na seção 3.3.2) para transformar a imagem de entrada em um conjunto de imagens binárias;

2. Computar, para cada imagem binária, a dimensão fractal correspondente aos contornos de regiões.

Para a fase de classificação é empregado um classificador *SVM* (*Support Vector Machine*) induzido por meio de um *kernel* polinomial empregando o algoritmo *SMO* (*Sequential Minimal Optimization*) [Platt 99]. O *SVM* foi escolhido por sua eficácia e por ser amplamente empregado na classificação de imagens médicas [Depeursinge 08, Unay 11], permitindo uma melhor comparação com outros trabalhos desenvolvidos na área.

4.2 Algoritmo de Extração de Características do *FFS*

O algoritmo de extração de características do *FFS* pode ser dividido em duas etapas principais. Na primeira delas é aplicada a técnica de decomposição em pilha de imagens binárias à imagem de entrada, resultando em um conjunto de imagens binárias. Quando uma imagem $I(x,y)$ é limiarizada por um valor $t, \in \{0,1,\cdots,n_t-1\}$, uma imagem binária correspondente é obtida:

$$I_b(x,y;t) = \begin{cases} 1 \text{ se } I(x,y) \geq t \\ 0, \text{ caso contrário.} \end{cases} \tag{4.1}$$

onde $I_b(x,y;t)$ denota a imagem binária obtida com o limiar t. Para uma imagem em níveis de cinza existem n_t possíveis imagens binárias a serem obtidas. A pilha de imagens binárias corresponde a esse conjunto de imagens.

A segunda etapa do algoritmo de extração *FFS* consiste em computar a dimensão fractal dos contornos das regiões de cada imagem binária. Os contornos de regiões de uma imagem binária $I_b(x,y;t)$ são computados de acordo com a equação 4.2:

$$\Delta(x,y;t) = \begin{cases} 1 \text{ se } \exists(x',y') \in N_8[(x,y)]: \\ \quad I_b(x',y';t) = 0 \ \wedge \\ \quad I_b(x,y;t) = 1, \\ 0, \text{ caso contrário.} \end{cases} \tag{4.2}$$

onde $N_8[(x,y)]$ é o conjunto de pixels que são 8-conectados a (x,y). $\Delta(x,y;t)$ toma o valor 1 se o pixel na posição (x,y) na imagem binária correspondente $I_b(x,y;t)$ tem o valor 1 e tem ao menos um pixel na 8-vizinhança com valor 0. Caso contrário, $I_b(x,y;t)$ toma o valor 0. Desta maneira, pode-se notar que os contornos resultantes terão espessura de um pixel.

A dimensão fractal $\mathcal{D}(t)$, onde t indica o valor de limiar usado para obter a imagem de contornos $\Delta(x,y;t)$, é computada por meio do algoritmo de contagem de caixas descrito na seção 3.3.3. O valor de $\mathcal{D}(t)$ descreve a complexidade de contorno de objetos que foram segmentados empregando o limiar t.

Ao variar o valor t é possível gerar uma curva $\mathcal{D}(t)$ vs. t. Essa curva é empregada como vetor de características para descrever a complexidade de contornos de estruturas e objetos segmentados por diferentes valores de limiar. Esse procedimento está ilustrado no diagrama da figura 4.1.

Existem dois motivos principais para se empregar a curva $\mathcal{D}(t)$ vs. t ao invés de um único valor de $\mathcal{D}(t)$ computado a partir de um único limiar para descrever uma imagem. O primeiro motivo se deve à dificuldade de se encontrar o limiar correto que separa estruturas e objetos do plano de fundo o que, em geral, é uma tarefa dependente do domínio de aplicação. O segundo motivo se deve ao fato de que, para determinadas imagens, uma segmentação binária pode não produzir um resultado satisfatório. Por exemplo, um único exame de imagem médica pode conter mais de duas estruturas anatômicas diferentes, cada uma das quais apresentando diferentes intervalos de valores de níveis de cinza. Esse exemplo será mais bem discutido na seção 4.3 na qual o método proposto é aplicado na tarefa de classificação de doenças pulmonares.

O algoritmo 4.1 sumariza o processo de extração do *FFS*, sendo que $\vec{v}_{FFS}$ denota o vetor de caracte-

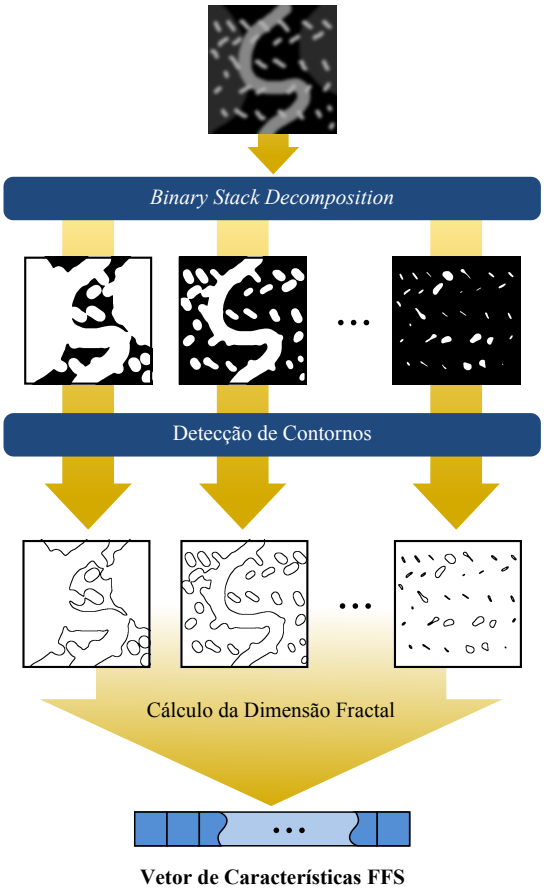


Figura 4.1: Extração de características pelo *FFS* tomando como entrada uma imagem artificial em níveis de cinza.

rísticas resultante. Na linha 2, T é o conjunto de todos os possíveis valores de níveis de cinza que uma imagem I pode assumir. O procedimento **Threshold** na linha 3 limiariza a imagem I utilizando t como valor de limiar como descrito na equação 4.1. O procedimento **FindBorders** na linha 4 corresponde à equação 4.2. A dimensão fractal dos contornos de regiões é calculada na linha 5 por meio do procedimento **BoxCounting**, que corresponde ao algoritmo de contagem de caixas, um dos mais conhecidos para o cálculo de dimensão fractal [Schroeder 92].

Algoritmo 4.1 Fast Fractal Stack (FFS).

Entrada: Imagem em níveis de cinza I.
Saída: Vetor de características $\vec{v}_{FFS}$.

1: $i \leftarrow 0$
2: **para** $t \in T \subseteq \{0, 1, \cdots, n_l\}$ **faça**
3: $\quad I_b(x,y;t) \leftarrow$ **Threshold**$(I,\ t)$
4: $\quad \Delta(x,y;t) \leftarrow$ **FindBorders**$(I_b(x,y;t))$
5: $\quad \mathcal{D}(t) \leftarrow$ **BoxCounting**$(\Delta(x,y;t))$
6: $\quad \vec{v}_{FFS}[i] \leftarrow \mathcal{D}(t), \quad i \leftarrow i+1$
7: **fim para**
8: **retorna** $\vec{v}_{FFS}$

É importante notar que a dimensão fractal pode ser eficientemente calculada em tempo compu-

tacional linear pelo algoritmo de contagem de caixas proposto em [Traina Jr. 00]. Desta maneira, a complexidade do algoritmo de extração do *FFS* é $O(N \cdot |T|)$, onde N é o número de pixels na imagem em níveis de cinza I e $|T|$ é o número de diferentes valores de limiar empregados para gerar a pilha de imagens binárias. Como será discutido na seção 4.2.1, T é apenas um pequeno subconjunto de todos os possíveis valores de limiar. Portanto, o algoritmo de extração do *FFS* tem custo linear com respeito ao tamanho da imagem.

4.2.1 Dimensionalidade do Vetor de Características

A dimensionalidade do vetor de características extraído pelo algoritmo *FFS* corresponde ao número de diferentes limiares empregados para gerar a pilha de imagens binárias. Ou seja, cada imagem binária contribui com um valor de $\mathcal{D}(t)$ para o vetor de características resultante. Se todos os n_l possíveis níveis de cinza fossem utilizados, o vetor de características resultante seria composto por n_l atributos. Por exemplo, para uma imagem na qual seus pixels podem tomar 256 diferentes valores de níveis de cinza, a dimensionalidade máxima de um vetor de características extraído seria também de 256.

De maneira intuitiva, pode-se concluir que utilizar todos os possíveis valores de limiar resultaria em um melhor desempenho de classificação, pois um maior número de características seriam extraídas, introduzindo mais informações ao processo de classificação. No entanto, isso não é verdade por duas razões. Primeiro, as imagens binárias obtidas por valores de limiar contíguos tendem a ser muito similares, resultando em valores de dimensão fractal altamente correlacionados e que não adicionam informações úteis ao processo de classificação. O segundo problema, conforme discutido na seção 4.1, deve-se ao fato do desempenho de classificação decair conforme o número de atributos aumenta devido à maldição da dimensionalidade.

Para tratar de ambos os problemas, é adotada a estratégia de selecionar um número fixo de limiares igualmente espaçados, coforme descrito na equação 4.3.

$$t_i = \left\lfloor \frac{n_l}{n_t + 1} \cdot i \right\rfloor, \;\; i = 1, 2, \cdots, n_t \tag{4.3}$$

onde n_t é o número de limiares a serem selecionados. Nos experimentos realizados, n_t foi empiricamente definido como oito. Apesar de simples, esta estratégia se mostrou eficaz na prática (conforme será demonstrado na seção 4.3), obtendo resultados que foram equivalentes ou melhores que escolher os atributos por meio de métodos de seleção supervisionados como o *CFS*. Adicionalmente, a abordagem adotada não requer qualquer conhecimento sobre a distribuição de classes do conjunto de imagens.

4.3 Experimentos

Em uma imagem de tomografia computadorizada, o nível de cinza das estruturas encontradas está relacionado à capacidade de se absorver o raio-X incidente. O ar, por exemplo, é menos denso que a água e, por este motivo, apresenta um menor valor de nível de cinza na imagem. Desta maneira, é possível identificar diferentes tecidos em uma imagem de tomografia dependendo do respectivo coeficiente de atenuação.

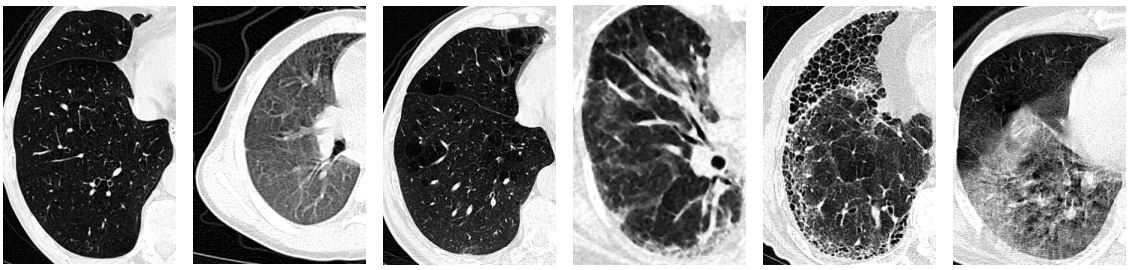
Nesta seção o método de extração *FFS* é avaliado na tarefa de classificar doenças pulmonares difusas (DPDs). O algoritmo de extração do *FFS* é utilizado para decompor a imagem de tomografia do pulmão em uma pilha de imagens binárias onde cada imagem binária corresponde a tecidos de diferentes

coeficientes de atenuação. A medida de complexidade de contornos de cada imagem binária é então empregada para predizer a ocorrência de DPDs que são caracterizadas por alterações no tecido pulmonar saudável.

Para avaliar o método de extração de características proposto, casos clínicos do período de 2001 a 2006 foram selecionados junto ao Hospital das Clínicas de Ribeirão Preto da Universidade de São Paulo. O conjunto é composto por 284 imagens de tomografias computadorizadas de 67 pacientes. Cada imagem possui 512 $\times$ 512 pixels e a espessura de cada fatia de tomografia é de 1mm. A profundidade de bits é 12 e foi convertida para 8 para o processo de extração de características.

A preparação da base de imagens consistiu em segmentar os pulmões do fundo em cada imagem de tomografia. Regiões de interesse (*regions of interest - ROIs*) contíguas de tamanho 64 $\times$ 64 pixels e sobreposição de 16 pixels entre duas *ROIs* adjacentes foram selecionadas a partir das regiões segmentadas como pulmões.

Cada *ROI* foi classificada por um especialista médico como normal ou um padrão de DPD. Os padrões de DPD mostrados na figura 4.2 foram os seguintes: (i) enfisema, (ii) consolidação, (iii) espessamento, (iv) favo de mel e (v) vidro fosco. A tabela 4.1 mostra a distribuição das classes para as *ROIs* selecionadas a partir da base de imagens de tomografia.



(a) Normal (b) Consolidação (c) Enfisema (d) Espessamento (e) Favo de Mel (f) Vidro Fosco

Figura 4.2: Exemplos de imagens de tomografia. (a) Normal, (b) consolidação, (c) enfisema, (d) espessamento, (e) favo de mel e (f) vidro fosco. Imagens provenientes do Hospital das Clínicas de Ribeirão Preto da Universidade de São Paulo.

Tabela 4.1: Distribuição das classes para as *ROIs* selecionadas da base de imagens de tomografias.

Classe	*ROIs*
Consolidação	451
Enfisema	502
Espessamento	590
Favo de Mel	530
Normal	590
Vidro Fosco	595

O restante desta seção é organizado da seguinte maneira. Na subseção 4.3.1 são apresentados os extratores de características empregados para comparar o desempenho do *FFS*. Por fim, na subseção 4.3.2 são apresentados os resultados obtidos nos experimentos.

4.3.1 Extratores de Características Utilizados para Comparação

Nos experimentos realizados para a tarefa de classificação de doenças pulmonares, o desempenho do *FFS* foi comparado com os seguintes extratores de características: histograma de níveis de cinza, descritores de Haralick, medidas da distribuição dos níveis de cinza e momentos de Zernike [Khotanzad 90]. Histogramas de níveis de cinza e descritores de Haralick são amplamente empregados em trabalhos com imagens de tomografias pulmonares [Uchiyama 03, Bugatti 09]. O histograma de níveis de cinza de uma imagem corresponde a função densidade de probabilidade de seus níveis de cinza, ou seja, para um dado nível de cinza, é retornado a freqüência com que o mesmo ocorre. Para a extração do histograma, as *ROIs* foram quantizadas para 16 níveis de cinza, resultando em um vetor de características com 16 componentes.

Como medida das texturas presentes nas *ROIs*, foram empregadas as sete primeiras sumarizações das matrizes de co-ocorrência de níveis de cinza propostas em [Haralick 79]. Tais sumarizações são denominadas de descritores de Haralick e para sua extração foram empregadas as matrizes de co-ocorrência calculadas para as distâncias 1, 2, 3, 4 e 5 e para as direções de 0°, 45°, 90°e 135°. O vetor de características resultante é composto por 140 componentes: 4 direções $\times$ 5 distâncias $\times$ 7 sumarizações.

Para descrever a distribuição dos níveis de cinza das imagens foram consideradas seis medidas: mediana, desvio padrão, obliqüidade (*skewness*) e contraste dos primeiros e segundos vizinhos. O contraste dos primeiros e segundos vizinhos corresponde à média de diferença dos níveis de cinza dos pixels distantes em uma e duas unidades entre si. Por fim, para extrair características de forma foram empregados os momentos de Zernike.

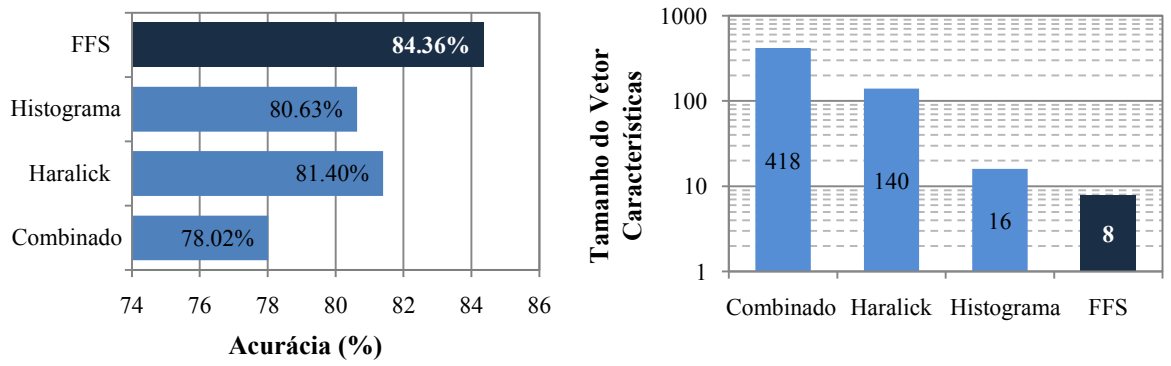
As características extraídas foram organizadas em três diferentes vetores de características. O primeiro corresponde às 16 componentes do histograma de níveis de cinza. O segundo corresponde aos 140 componentes dos descritores de Haralick. Por fim, foi utilizado um último vetor de características referido pelo termo "combinado" que corresponde a todas as características descritas nessa seção, ou seja, histograma, descritores de Haralick, distribuição dos níveis de cinza e momentos de Zernike. Todos os vetores de características tiveram seus componentes normalizados no intervalo $(0, 1)$.

4.3.2 Resultados dos Experimentos

Nesta seção são apresentados os resultados dos experimentos realizados. Para a etapa de classificação foi utilizado um classificador *SVM* com *kernel* polinomial empregando o algoritmo *SMO*. Os melhores parâmetros para o classificador *SVM* foram encontrados por meio de validação cruzada com 10 partições.

A figura 4.3 (a) mostra a acurácia de classificação obtida utilizando cada um dos métodos de extração. Os resultados foram obtidos realizando dez repetições de validação cruzada com dez partições. O *FFS* obteve uma acurácia média de 84,4%, superando os outros métodos de extração.

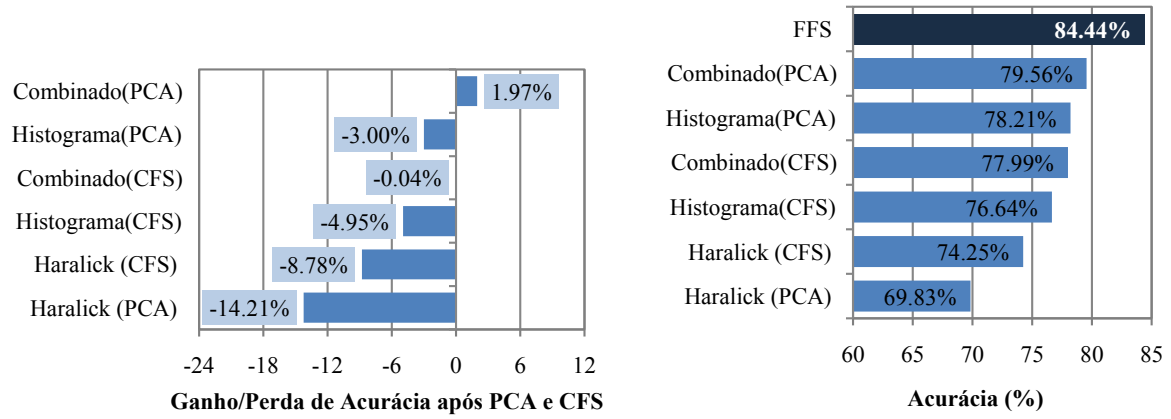
Adicionalmente, conforme mostrado na figura 4.3(b), o *FFS* apresentou a vantagem de obter um vetor de características com um menor número de componentes quando comparado com os demais métodos de extração. Este resultado é importante quando se considera o problema da maldição da dimensionalidade. A figura 4.4 exibe os resultados obtidos quando se aplicam os métodos *Principal Component Analysis* (*PCA*) e *Correlation Based Feature Selection* (*CFS*) para reduzir o número de componentes dos vetores de histograma, Haralick e combinado para 8 atributos, que corresponde ao mesmo número de atributos do vetor de características do *FFS*. A figura 4.4 (a) mostra o ganho e perda de acurácia após se empregar



(a) (b)

Figura 4.3: (a) Acurácia sem seleção de atributos para o método proposto (*FFS*), histograma, Haralick e vetor de características combinado. (b) Número de componentes dos vetores de características para cada método de extração.

o *PCA* e o *CFS*. Os resultados da figura 4.4(b) mostram que o *FFS* foi capaz de obter acurácia superior aos outros métodos. Isto indica que o *FFS* é capaz de obter uma representação mais compacta e com alto poder de descrição dos padrões de DPDs.



(a) (b)

Figura 4.4: Ganho e perda de acurácia após aplicação do *PCA* e *CFS* ao vetores de histograma, Haralick e combinado. (b) Comparação da acurácia com o *FFS*.

Foi também investigado o desempenho na tarefa de detecção de *ROIs* não-normais (com presença de algum padrão de doença pulmonar). *ROIs* classificadas como DPD foram consideradas casos positivos e *ROIs* normais foram consideradas casos negativos. Ao variar o limiar de classificação positivo do classificador *SVM*, curvas *ROC* (*Receiver Operating Characteristics*) foram geradas como gráficos da taxa de verdadeiros positivos (TVP) vs. taxa de falsos positivos (TFP). A figura 4.5 (a) mostra as curvas *ROC* para os métodos de extração. Uma vez que a parte superior esquerda (TVP = 1, TFP = 0) do espaço *ROC* corresponde ao ponto ótimo de operação do classificador, a figura 4.5 (b) mostra que o *FFS* obteve uma melhor performance de classificação em comparação aos demais extratores.

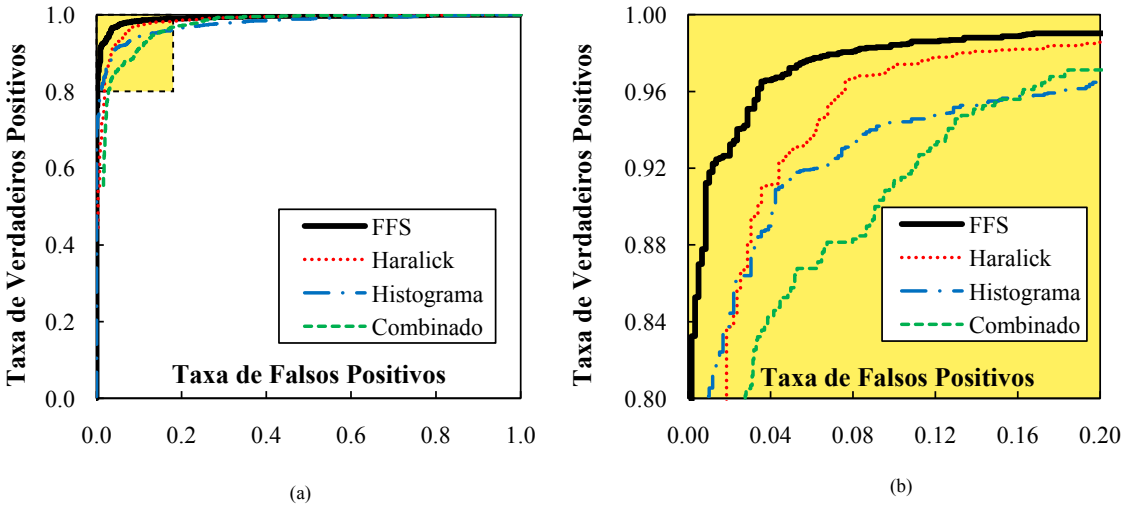


Figura 4.5: (a) Curva *ROC* para detecção de doenças pulmonares difusas. (b) Zoom da curva (a) no ponto de operação ótimo do classificador no espaço *ROC*.

4.4 Conclusões

Nesta seção foi apresentado o novo método de extração de características desenvolvido durante este projeto de Mestrado, o *Fast Fractal Stack* (*FFS*). O *FFS* emprega análise fractal para medir a complexidade dos contornos de estruturas e objetos presentes em imagens retornando um vetor de características compacto e com alto poder descritivo.

O *FFS* foi avaliado para a tarefa de classificação de doenças pulmonares difusas (DPDs) em imagens de tomografia do pulmão e obteve acurácia superior a 84% sem empregar seleção de atributos ou transformação de características. Os resultados demonstraram a eficácia do *FFS* em detectar e classificar cinco diferentes tipos de padrões de doenças pulmonares.

Adicionalmente, o *FFS* apresenta um algoritmo de extração eficiente. Enquanto a maioria dos métodos de extração são ao menos quadráticos, o custo computacional do algoritmo proposto é linear com respeito ao tamanho da imagem (número de pixels). Por este motivo, o *FFS* é uma solução promissora para sistemas de recuperação de imagens por conteúdo apoiando processos de decisão interativos.

Capítulo 5

O Método *Concept*

Neste capítulo é descrito um novo método de classificação de imagens proposto durante este projeto de Mestrado. O método proposto, denominado *Concept*, introduz aprimoramentos significativos na classificação de imagens por regras de associação. Seu desenvolvimento teve como base o método *IDEA* [Ribeiro 08] que também emprega regras de associação para realizar a classificação de imagens. O principal desses aprimoramentos refere-se ao modo como as representações das imagens são obtidas. No *IDEA* e nos demais métodos de classificação associativa, os vetores de características extraídos das imagens são discretizados para obter *itemsets* que irão representar as imagens no processo de classificação. Assim, cada item que compõe a representação da imagem corresponde a um intervalo de discretização de um atributo numérico. O problema desta abordagem é que uma grande quantidade de itens pode ser gerada, muitos deles sendo irrelevantes ou contendo pouca informação útil para o processo de classificação. Uma vez que o custo computacional de se minerar regras de associação depende da quantidade de itens existente na base, o processo de classificação por regras de associação pode ficar computacionalmente ineficiente e ter sua acurácia e precisão reduzidas.

Para tratar deste problema, o método *Concept* utiliza um novo algoritmo também desenvolvido neste projeto de mestrado e denominado *MFS-Map* (*Multi Feature Space Map*). O *MFS-Map* emprega análise de agrupamentos em diferentes espaços de características para obter a representação das imagens na forma de *itemsets*. Sua principal vantagem está no fato de que os itens obtidos são capazes de aproveitar de maneira bastante eficiente as informações contidas nas características extraídas das imagens para o processo de classificação. Isto é possível porque cada um dos itens corresponde a regiões dos espaços de características nas quais as imagens são visualmente similares. Nos experimentos realizados para a tarefa de classificação de imagens médicas o desempenho do *Concept* se mostrou superior ao do método *IDEA*, indicando que a abordagem de representação é promissora para a classificação associativa de imagens.

5.1 Descrição do Algoritmo *Concept*

Concept é um algoritmo de indução de classificadores associativos de imagens. O algoritmo toma como entrada um conjunto de imagens de treinamento $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ e as classes associadas a cada uma

das imagens $\mathbf{C} = \{c_1, c_2, \ldots, c_N\}$. A partir de $\mathbf{I}$ e $\mathbf{C}$ é retornado um conjunto de regras de associação $\mathbf{S}$ que será utilizado pelo classificador *Concept* para sugerir palavras chaves para uma imagem.

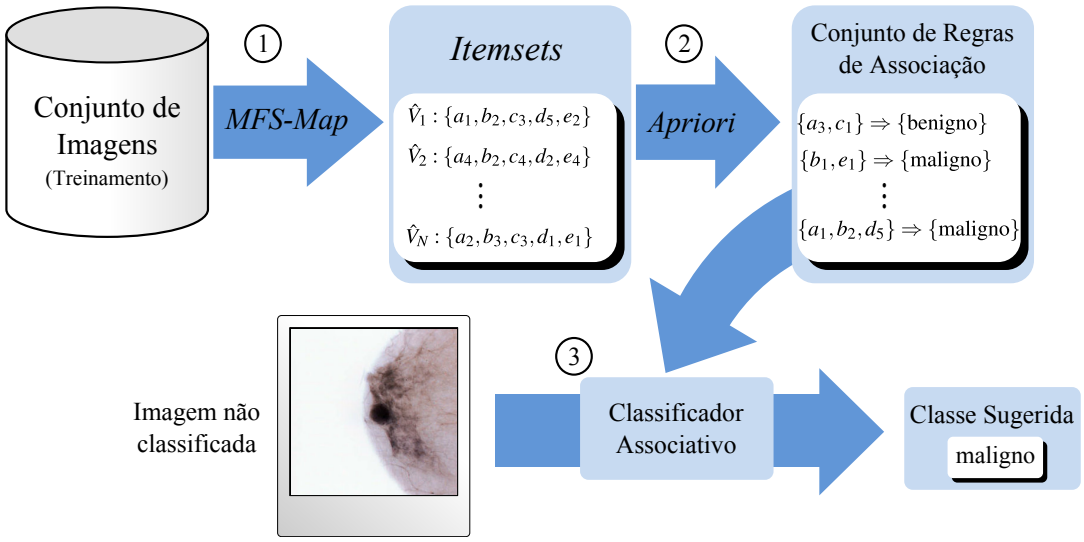


Figura 5.1: Diagrama do algoritmo *Concept*. Primeiramente a base de imagens é mapeada para um conjunto de *itemsets* pelo algoritmo *MFS-Map*. Em seguida, o algoritmo *Apriori* é utilizado para minerar regras de associação. Por fim, o classificador associativo do algoritmo *Concept* retorna a classe sugerida para a imagem.

O diagrama da figura 5.1 ilustra o processo de mineração de regras de associação do algoritmo *Concept* e classificação de imagens. A primeira etapa consiste em obter uma representação da base de imagens adequada para mineração de regras de associação. Para este fim é empregado o algoritmo *MFS-Map* (seção 5.2). A saída do algoritmo é a representação da base de imagens na forma transacional, na qual cada imagem I_i é representada por um *itemset*. Cada um dos *itemsets* é denotado por $\hat{V}_i$ e é denominado vetor de características transacional. O conjunto de todas as transações $\hat{\mathbf{V}}$ é então submetido ao algoritmo de mineração de regras de associação *Apriori* (discutido na seção 2.1.2) adotando-se duas restrições quanto às regras que podem ser geradas: o antecedente é composto somente por itens retornados pelo algoritmo *MFS-Map* e o conseqüente somente pelo valor do atributo classe. A equação 5.1 exemplifica uma possível regra de associação minerada pelo classificador *Concept* para um conjunto de imagens de mamografia.

$$\{a_1, b_2, d_5\} \Rightarrow \{\text{maligno}\} \tag{5.1}$$

A regra da equação 5.1 pode ser interpretada da seguinte maneira: imagens para as quais sua representação transacional contém os itens $\{a_1, b_2, d_5\}$ tendem a ser diagnosticadas com a palavra chave *maligno*. O algoritmo 5.1 descreve o processo de mineração de regras de associação. Na linha 1 é calculado o modelo de mapeamento denotado por M' por meio do procedimento **MfsMapCompute**. Nas linhas 2-6 o modelo de mapeamento é utilizado para mapear o conjunto de imagens para a forma transacional por meio do procedimento **MfsMap**. Os procedimentos **MfsMapCompute** e **MfsMap** fazem parte do algoritmo *MFS-Map*. Nas linhas 7-10 o valor do atributo classe é agrupado com a representação transacional de cada imagem para gerar seu respectivo *itemset*. Por fim, na linha 11 é realizada a mineração de regras de associação utilizando o algoritmo *Apriori* e na linha 12 são retornadas

as regras mineradas e o modelo de mapeamento.

Algoritmo 5.1 Mineração de regras de associação pelo algortimo *Concept*.

Entrada: Conjunto de imagens de treinamento $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ e valores do atributo classe $\mathbf{C} = \{c_1, c_2, \ldots, c_N\}$.
Saída: Conjunto de regras de associação $\mathbf{S}$ e modelo de mapeamento M'.

1: $M' \leftarrow$ **MfsMapCompute**$(\mathbf{I}, \mathbf{C})$ // *Calcula o modelo de mapeamento.*
2: $\hat{\mathbf{V}} \leftarrow \emptyset$
3: **para** $I_i \in \mathbf{I}$ **faça**
4: $\quad \hat{V}_i \leftarrow$ **MfsMap**(I_i, M') // *Converte a imagem para o formato transacional.*
5: $\quad \hat{\mathbf{V}} \leftarrow \hat{\mathbf{V}} \cup \hat{V}_i$
6: **fim para**
7: **para** $i \in \{1..N\}$ **faça**
8: $\quad t_i \leftarrow \{c_i \cup \hat{V}_i | c_i \in \mathbf{C} \wedge \hat{V}_i \in \hat{\mathbf{V}}\}$ // *Agrupa c_i à representação transacional $\hat{V}$.*
9: $\quad T \leftarrow T \cup \{t_i\}$ // *Adiciona transação t_i a T.*
10: **fim para**
11: $\mathbf{S} \leftarrow$ **Apriori**(T)
12: **retorna** $\mathbf{S}, M'$

Para realizar a classificação de uma nova imagem (etapa três no diagrama da figura 5.1), o algoritmo *Concept* toma como entrada a imagem a ser classificada, o conjunto de regras de associação $\mathbf{S}$ e o modelo de mapeamento M' obtido como saída do processo de mineração descrito no algoritmo 5.1 e retorna como saída todos os possíveis valores de atributo classe juntamente com um valor de *escore de confiança*. O valor de atributo classe que apresentar o maior valor de *escore de confiança* é aquele utilizado para classificar a imagem.

O *escore de confiança* de um valor de atributo classe c' corresponde à média de confiança das regras que apresentaram c' no conseqüente e que foram aplicáveis à imagem sendo classificada. Para calcular esta medida o algoritmo *Concept* emprega o conceito de casamento entre uma regra de associação e a representação transacional da imagem a ser classificada, o que corresponde a verificar se a regra é aplicável à imagem. A representação transacional *satisfaz* uma regra quando contém todos os itens presentes do antecedente da regra. A equação 5.2 ilustra este conceito:

$$\hat{V} : \{a_1, b_3, c_2\} \xrightarrow{\text{satisfaz}} \{b_3, c_2\} \Rightarrow \{\text{benigno}\} \tag{5.2}$$

Uma vez que $\hat{V} \supseteq \{b_3, c_2\}$, ou seja, $\hat{V}$ contém os itens b_3 e c_2, pode-se dizer que a regra da equação 5.2 é satisfeita. Outro exemplo é dado na equação 5.3:

$$\hat{V} : \{a_1, b_3, c_2\} \xrightarrow{\text{não satisfaz}} \{a_1, b_2\} \Rightarrow \{\text{maligno}\} \tag{5.3}$$

Uma vez que $\hat{V} \not\supseteq \{a_1, b_2\}$ pois $\hat{V}$ não contém o item b_2, então $\hat{V}$ não satisfaz a regra da equação 5.3.

O algoritmo 5.2 descreve a predição de classe pelo classificador *Concept*. Na linha 1, a representação transacional $\hat{V}_I$ da imagem a ser classificada é extraida pelo algoritmo *MFS-Map* por meio do mapeamento retornado pelo algoritmo 5.1. A linha 2 consiste em encontrar todas as regras que a representação transacional da imagem satisfaz e atribuí-las ao conjunto $\mathbf{S}'$. Na linha 3 calcula-se a variável minConf que corresponde ao máximo valor de confiança dentre todas as regras que $\hat{V}_I$ pode satisfazer menos o parâmetro τ. Desta maneira, minConf é um limiar mínimo de confiança para as regras a serem

consideradas durante a classificação. Regras que apresentarem confiança menor que minConf serão descartadas. O valor de τ foi definido empiricamente nos experimentos realizados como 0,1.

Nas linhas 4-6 as regras que apresentam confiança maior que τ são divididas em subconjuntos de acordo com o valor do atributo classe encontrado em seu consequente. Por exemplo, $\mathbf{S}'_1$ irá conter todas as regras que apresentarem confiança maior que τ e que apresentarem em seu consequente o valor de classe c'_1.

Para cada um dos possíveis valores de classe $c'_i \in \mathcal{C} = \{c'_1, \cdots, c'_m\}$ é calculado um valor de $escore_i$ que corresponde a média das confianças das regras contidas em $\mathbf{S}'_i$. Por fim, na linha 10 é retornado o valor de classe com maior valor de *escore*.

Algoritmo 5.2 Predição de classe pelo algoritmo *Concept*.

Entrada: Imagem I a ser classificada, conjunto de regras de associação $\mathbf{S}$ e modelo de mapeamento M'.
Saída: Valor de atributo classe c'_i e valor de escore de confiança $escore_i$ associado.

1: $\hat{V}_I \leftarrow \mathbf{MfsMap}(I, M')$
2: $\mathbf{S}' \leftarrow \{r \mid \hat{V}_I \text{ satisfaz } r \in \mathbf{S}\}$
3: $\text{minConf} \leftarrow \max\{\text{conf}(r) - \tau,\ r \in \mathbf{S}'\}$
4: **para** cada valor de classe $c'_i \in \mathcal{C} = \{c'_1, \cdots, c'_m\}$ **faça**
5: $\quad \mathbf{S}' \leftarrow \{r : X \Rightarrow \{c\} \mid c = c'_i \wedge \text{conf}(r) > \text{minConf},\ r \in \mathbf{S}'\}$
6: **fim para**
7: **para** $\mathbf{S}'_i \in \{\mathbf{S}'_1, \mathbf{S}'_2, \cdots, \mathbf{S}'_m\}$ **faça**
8: $\quad escore_i \leftarrow 1/|\mathbf{S}'_i| \sum_{r \in \mathbf{S}'_i} \text{conf}(r)$
9: **fim para**
10: **retorna** $c'_i,\ escore_i = \max\{escore_1, \cdots, escore_m\}$

5.2 O Algoritmo *MFS-Map*

Para aplicar um classificador associativo sobre um conjunto de imagens é preciso mapeá-lo para a forma transacional, na qual cada imagem corresponde a um *itemset*. Conforme discutido na seção 2.1.3, os métodos de classificação associativa existentes na literatura fazem este mapeamento por meio da discretização dos vetores de características extraídos das imagens, convertendo cada um dos intervalos de discretização para um item que irá compor um *itemset*. O problema de se empregar discretização de atributos é que uma grande quantidade de intervalos de discretização podem ser gerados resultando em itens irrelevantes. Uma vez que o custo computacional de se minerar regras de associação depende da quantidade de itens existentes na base, o processo de classificação por regras de associação pode ficar extremamente custoso computacionalmente e ter sua acurácia e precisão reduzidas.

O algoritmo *MFS-Map* (*Multi Feature Space Map*) apresentado nesta seção e desenvolvido durante este projeto de mestrado adota uma nova abordagem para mapear o conjunto de imagens para a forma transacional. Inicialmente são aplicados diferentes extratores para cada imagem da base. O resultado deste processo é um conjunto de vetores de características para cada imagem. Cada um desses vetores corresponde a um ponto em um espaço multidimensional, sendo que o número de espaços corresponde ao número de extratores empregados.

A segunda etapa do processo consiste em encontrar centróides de agrupamentos em cada um dos espaços. Assim, os centróides correspondem a regiões do espaço nas quais as imagens são visualmente

similares. Desta maneira, a representação da imagem consiste no conjunto de rótulos dos centróides mais próximos da representação vetorial da imagem em cada um dos espaços de características. No *MFS-Map* o número de itens que compõem a representação transacional é igual ao número de extratores empregados. Considere o exemplo 5.2.1:

Exemplo 5.2.1. *De uma base de imagens médicas de tomografias do pulmão sobre a qual se deseja aplicar um algoritmo de classificação são extraídos os seguintes vetores de características:*

1. *Histograma de níveis de cinza (16 componentes);*
2. *Descritores de Haralick (140 componentes);*
3. *Vetor de características do* FFS*, descrito no capítulo 4 (8 componentes).*

Para realizar a classificação utilizando um algoritmo de classificação associativa tradicional, os três vetores de características poderiam ser organizados em um único vetor de 164 componentes. Em seguida, um algoritmo de discretização é aplicado sobre cada um dos componentes do vetor. O *itemset* que irá representar a imagem consiste nos intervalos de discretização para cada um dos atributos.

Já para o algoritmo *Concept* os três vetores são tratados independentemente. Ou seja, inicialmente são encontrados centróides de agrupamento para o espaço dos vetores de características do histograma de níveis de cinza e então esse processo é repetido para os descritores de Haralick e para o extrator *FFS*. Para obter a representação por *itemset* de uma imagem são extraídos os três vetores de características da imagem. O primeiro item que irá compor a representação da imagem corresponde ao centróide mais próximo do vetor de histograma da imagem no espaço dos vetores de histograma. O segundo e terceiro item correspondem, respectivamente, aos centróides mais próximos do vetor de características da imagem no espaço de vetores de características de Haralick e *FFS*. Desta maneira, pode-se notar que o número de itens na representação transacional da imagem é igual ao número de extratores empregados (neste caso, três).

A vantagem do algoritmo *MFS-Map* está no fato de que a representação obtida carrega informações semânticas valiosas para o processo de classificação, diferente dos métodos de classificação associativa tradicionais nos quais os itens representam somente intervalos de discretização que tendem a carregar menos informações semânticas. Adicionalmente, a quantidade de itens gerados pelo *MFS-Map* é significativamente menor quando comparada com as abordagens baseadas em discretização de atributos, tornando o processo de mineração de regras de associação mais eficiente.

O algoritmo *MFS-Map* pode ser dividido em duas etapas principais: (i) cálculo do modelo de mapeamento e (ii) mapeamento da base de imagens para representações transacionais. Para calcular o modelo de mapeamento, denotado por M', o *MFS-Map* toma como entrada um conjunto de imagens $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$. O modelo de mapeamento é então utilizado para transformar o conjunto de imagens $\mathbf{I}$ em um conjunto de transações $\hat{\mathbf{V}} = \{\hat{V}_1, \hat{V}_2 \cdots, \hat{V}_N\}$ onde $\hat{V}_i$ correponde à representação transacional da i-ésima imagem I_i.

Para realizar o cálculo do modelo de mapeamento, o conjunto de imagens é mapeado em K espaços de características $\{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_K\}$ por meio de K diferentes funções de extração denotadas por $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_K\}$. Em cada um dos espaços de características $\mathcal{F}_i$ é aplicado o algoritmo de agrupamento descrito na seção 5.3 resultando em um conjunto de centróides de agrupamentos $G_i = \{g_{i,1}, g_{i,2}, \cdots, g_{i,n}\}$.

A figura 5.2 ilustra o processo do cálculo de agrupamento para o caso no qual são utilizados três espaços de características. Na primeira etapa, os extratores ε_1, ε_2 e ε_3 são utilizados para mapear o conjunto de imagens em três espaços vetoriais: $\mathcal{F}_1$, $\mathcal{F}_2$ e $\mathcal{F}_3$. Mapear o conjunto de imagens utilizando um extrator, neste contexto, significa converter cada uma das imagens em um vetor que corresponde a um ponto em seu respectivo espaço de características. Na figura 5.2 os espaços de características estão representados como espaços tridimensionais, sendo que cada ponto corresponde a uma imagem. É importante notar que no caso geral os espaços de características não precisam ser tridimensionais e nem mesmo ter a mesma dimensionalidade entre si.

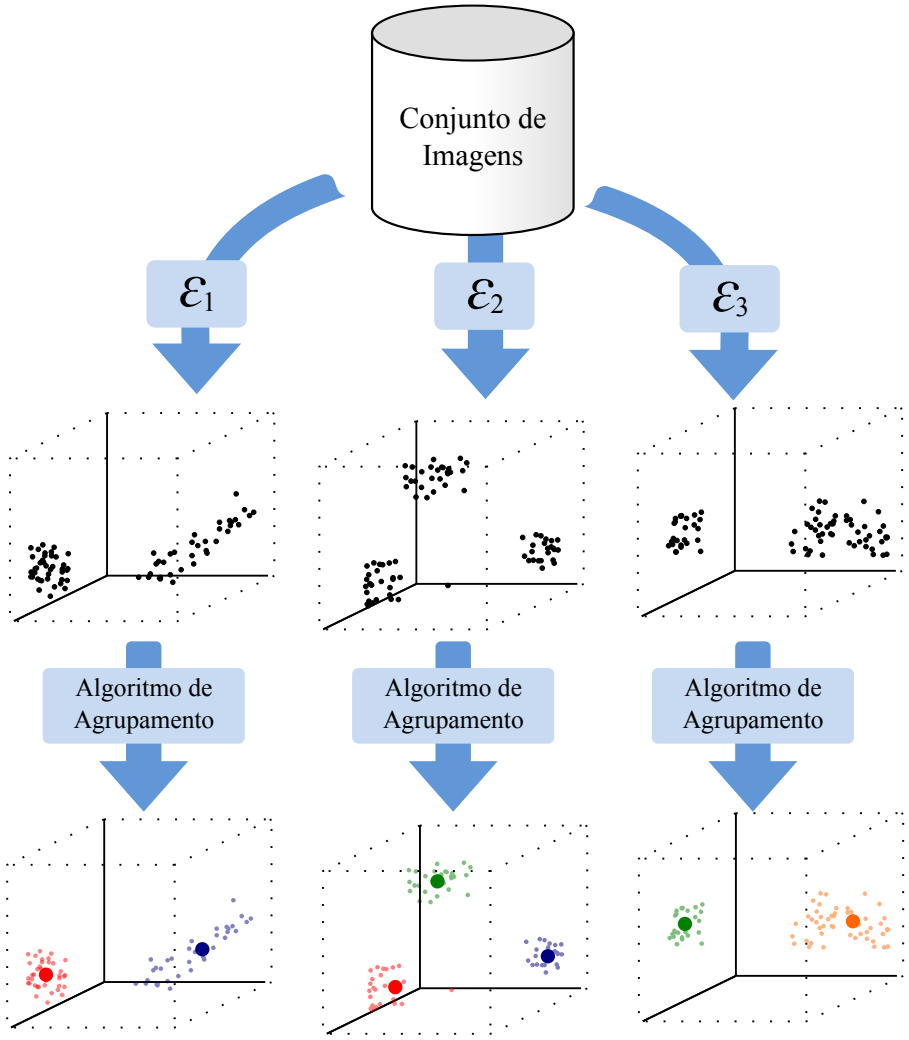


Figura 5.2: Diagrama do algoritmo *MFS-Map*. Inicialmente o conjunto de imagens é mapeado para diferentes espaços de características. Em seguida, o modelo de mapeamento é calculado como os centróides encontrados em cada um dos espaços de características por meio da aplicação de um algoritmo de agrupamento.

Na segunda etapa do diagrama da figura 5.2 é aplicado o algoritmo descrito na seção 5.3 para encontrar os centróides dos agrupamentos em cada um dos espaços de características. O j-ésimo centróide do i-ésimo espaço de características é denotado por $g_{i,j}$. O modelo de mapeamento corresponde ao conjunto de centróides em cada um dos espaços de características, ou seja, $M' = \{G_1, G_2, \cdots, G_K\}$.

O algoritmo 5.3 (**MfsMapCompute**) descreve o cálculo do modelo de mapeamento. Nas linhas 1-3

é aplicado cada um dos extratores ao conjunto de imagens resultando em um espaço de características. Nas linhas 5-8 é aplicado o algoritmo de agrupamento sobre o conjunto de imagens mapeado em cada um dos espaços de características $\mathcal{F}_i$. A saída do algoritmo de agrupamento é um conjunto de centróides G_i para cada um dos espaços de características (linha 6). Por fim, na linha 9, o modelo de mapeamento denotado por M' é retornado.

Algoritmo 5.3 MfsMapCompute: cálculo do modelo de mapeamento do *MFS-Map*.

Entrada: Conjunto de imagens de treinamento $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ e valores do atributo classe $\mathbf{C} = \{c_1, c_2, \ldots, c_N\}$.

Saída: Modelo de mapeamento M'.

1: **para** cada extrator $\varepsilon_i \in E = \{\varepsilon_1, \cdots, \varepsilon_K\}$ **faça**
2: $\quad \mathcal{F}_i \leftarrow \varepsilon_i(\mathbf{I})$
3: **fim para**
4: $M' \leftarrow \emptyset$
5: **para** cada $\mathcal{F}_i \in \{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_K\}$ **faça**
6: $\quad G_i \leftarrow$ **EncontraCentróides**$(\mathcal{F}_i, \mathbf{C})$
7: $\quad M' \leftarrow M' \cup G_i$
8: **fim para**
9: **retorna** M'

Uma vez calculado o modelo de mapeamento M', o processo de obtenção da representação transacional de uma imagem consiste em, primeiramente, extrair representações vetoriais da imagem utilizando o mesmo conjunto de extratores $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_K\}$ utilizados para calcular M'. Como resultado, a imagem irá corresponder a um ponto em cada um dos espaços de características $\{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_K\}$. Para cada espaço de características $\mathcal{F}_i$ é então encontrado o centróide mais próximo do vetor de características utilizando a função de distância Euclidiana. A representação transacional da imagem consiste no conjunto de rótulos de centróides mais próximos da representação vetorial $\vec{v}_i$ da imagem em cada um dos espaços de características. Desta maneira, cada um dos itens corresponde a regiões do espaço de características nas quais as imagens são visualmente similares. A figura 5.3 ilustra a obtenção da representação transacional de uma imagem utilizando o mesmo mapeamento que foi calculado no diagrama da figura 5.2. Calculando os centróides mais próximos da representação da imagem em cada um dos espaços de características têm-se como resultado o *itemset* $\hat{V} = \{g_{1,2}, g_{2,1}, g_{3,1}\}$.

O algoritmo 5.4 (**MfsMap**) descreve como é realizado o mapeamento de uma imagem utlizando o modelo de mapeamento M' calculado no algoritmo 5.3. Nas linhas 1-3 são extraídos os vetores de características da imagem. É importante notar que os extratores sendo utilizados são os mesmos que foram empregados para calular M'. Nas linhas 4-8 o vetor de características transacional da imagem é calculado. Para tanto, na linha 6 é encontrado o centróide mais próximo da representação vetorial da imagem em cada um dos espaços de características. O centróide é então inserido no vetor de caterísticas transacional que é retornado na linha 9. Uma característica relevante do *MFS-Map* é que as representações transacionais das imagens correspondem a *K-itemsets*, ou seja, *itemsets* nos quais o número de elementos é igual ao número de espaços de características utilizados para mapear o conjunto de imagens.

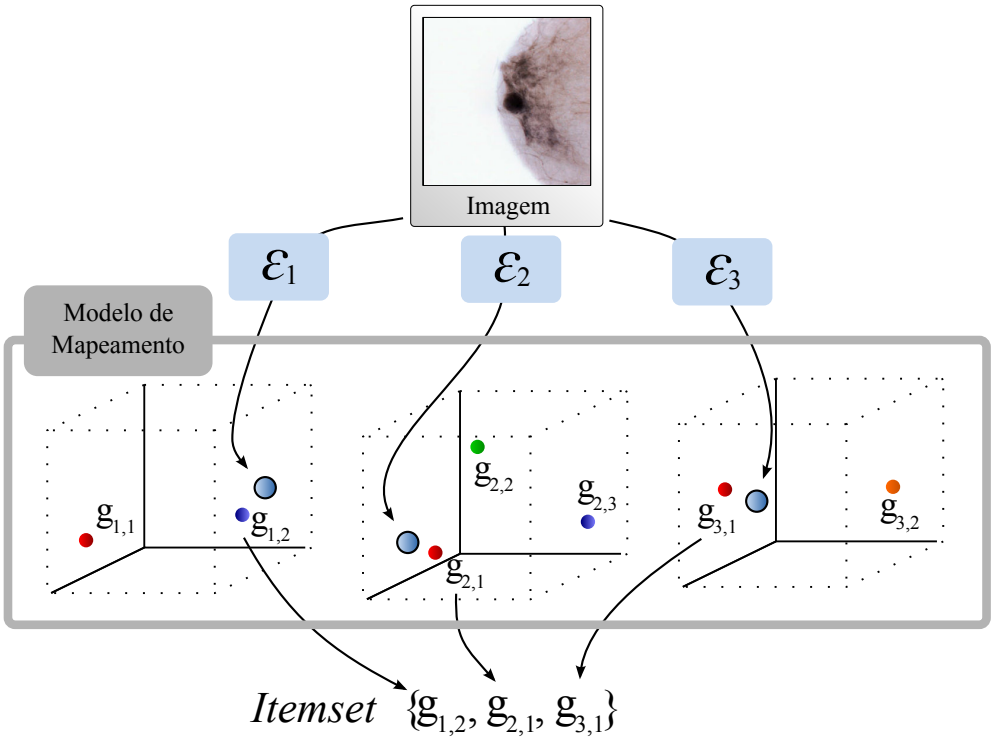


Figura 5.3: Mapeamento de uma imagem para um *itemset* pelo algoritmo *MFS-Map*.

Algoritmo 5.4 MfsMap: mapeamento de uma imagem para sua representação transacional.

Entrada: Imagem I e modelo de mapeamento M'.
Saída: Representação transacional $\hat{V}$.

1: **para** cada extrator $\varepsilon_i \in E = \{\varepsilon_1, \cdots, \varepsilon_K\}$ **faça**
2: $\quad \vec{v}_i \leftarrow \varepsilon_i(I)$
3: **fim para**
4: $\hat{V} \leftarrow \emptyset$
5: **para** cada vetor de características $\vec{v}_i \in \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_K\}$ **faça**
6: $\quad g_{i,j} \leftarrow$ **CentróideMaisPróximo**$(\vec{v}_i, M')$
7: $\quad \hat{V} \leftarrow \hat{V} \cup g_{i,j}$
8: **fim para**
9: **retorna** $\hat{V}$

5.3 Algoritmo de Agrupamento do *MFS-Map*

No algoritmo *MFS-Map* os itens que irão compor a representação transacional da imagem correspondem aos rótulos dos centróides que estão mais próximos da representação vetorial da imagem em cada um dos espaços de características. Nesta seção é descrito o algoritmo de agrupamento utilizado para encontrar as coordenadas destes centróides.

O algoritmo de agrupamento do *MFS-Map* possui duas entradas. A primeira delas consiste no conjunto de vetores de características denotados por $\mathbf{V} = \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_N\}$ que correspondem às representações das imagens em um espaço de características $\mathcal{F}$ e são obtidas por meio de uma função de extração ε. Desta maneira, cada vetor de característica pode ser visto como um ponto em um espaço n-dimensional, onde n é a dimensão do espaço de características e também o número de componentes de cada vetor. Adicionalmente, define-se uma função de distância d entre esses pontos que pode ser

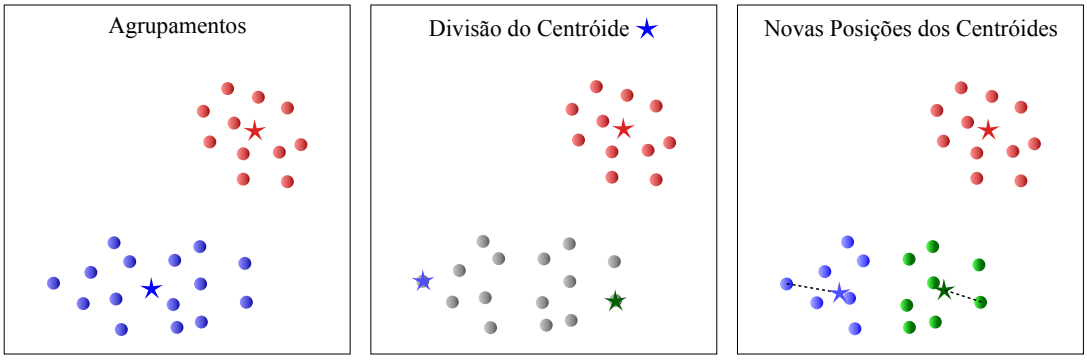
interpretada como uma medida de dissimilaridade. A segunda entrada do algoritmo é o conjunto de valores de classe $\mathbf{C} = \{c_1, c_2, \cdots, c_N\}$ para cada um dos vetores de características. Desta maneira, a classe associada ao i-ésimo vetor de características $\vec{v}_i$ é denotada por c_i.

A saída do algoritmo é o conjunto de centróides que é calculado por meio de uma adaptação do algoritmo *x-Means*, descrito na seção 2.2. O objetivo é separar os vetores de características em agrupamentos denotados por θ. Cada agrupamento possui um centróide g que corresponde a um ponto n-dimensional, onde n é a dimensionalidade do espaço de características. Os vetores de características são associados ao agrupamento que possui o centróide mais próximo.

No algoritmo *x-Means*, para avaliar a qualidade dos agrupamentos e verificar se um centróide deve ser dividido ou não, é empregado o critério de informação bayesiano (*Bayesian Information Criterion - BIC*). Ao empregar o *BIC* como métrica de qualidade no algoritmo *x-Means* é adotada a suposição de que os agrupamentos podem ser modelados por funções gaussianas esféricas no espaço de características.

O algoritmo de agrupamentos do *MFS-Map* consiste em, inicialmente, executar o algoritmo *k-Means* com um valor inicial mínimo k_0 de agrupamentos. Cada um dos agrupamentos obtidos é avaliado com respeito à medida $\mathcal{H}$ (ver equação 5.4). Os centróides dos agrupamentos que apresentaram os maiores valores para a medida $\mathcal{H}$, que pode ser interpretada como uma medida de impureza, são divididos em dois novos centróides. Os novos centróides correspondem a dois pontos associados ao agrupamento escolhidos aleatoriamente. O objetivo deste procedimento é preservar os agrupamentos satisfatórios e recalcular os demais agrupamentos. O algoritmo é executado iterativamente enquanto o número de agrupamentos é menor que um valor máximo de agrupamentos denotado por $k_{\max}$. A cada iteração são armazenados os agrupamentos obtidos e também uma medida de qualidade para o conjunto de agrupamentos que é denotada por $\mathcal{Q}$ (ver equação 5.5). Ao fim das iterações é retornado o conjunto de centróides de agrupamentos que apresentou o maior valor para a medida de qualidade $\mathcal{Q}$.

A figura 5.4 ilustra uma iteração do algoritmo de agrupamentos do *MFS-Map*. Na figura 5.4(a) são mostrados dois agrupamentos obtidos após a execução do algoritmo *k-Means* sendo que os respectivos centróides estão marcados com o símbolo $\star$. Na figura 5.4(b) o centróide localizado na parte inferior é dividido. Para tanto, dois pontos pertencentes ao respectivo agrupamento são escolhidos aleatoriamente para serem os novos centróides. Na figura 5.4(c) cada um dos pontos é atribuído ao centróide mais próximo e a posição dos centróides é recalculada.



(a) (b) (c)

Figura 5.4: Exemplo de uma iteração do algoritmo de agrupamento do *MFS-Map*.

O algoritmo 5.5 (**EncontraCentróides**) descreve o algoritmo proposto. Nas linhas 1-2 os vetores de

características $\mathbf{V}$ e os respectivos valores de classe $\mathbf{C}$ são particionados em dois subconjuntos disjuntos: $\mathbf{V}^A \cap \mathbf{V}^B = \emptyset$ e $\mathbf{V}^A \cup \mathbf{V}^B = \mathbf{V}$. Os valores de classe associados aos vetores de características de $\mathbf{V}^A$ e $\mathbf{V}^B$ são, respectivamente, $\mathbf{C}^A$ e $\mathbf{C}^B$. O objetivo de separar os vetores de características em dois subconjuntos é usar $\mathbf{V}^A$ para calcular as posições dos centróides e $\mathbf{V}^B$ para avaliar a qualidade dos centróides obtidos. Adicionalmente, é importante notar que o procedimento **AmostragemEstratificada** na linha 1 realiza uma amostragem estratificada, ou seja, o subconjunto retornado apresenta a mesma distribuição de classes do conjunto original.

Algoritmo 5.5 EncontraCentróides: cálculo de centróides de agrupamento.

Entrada: Conjunto de vetores de características $\mathbf{V}$ e valores de atributo classe associados $\mathbf{C}$.
Saída: Conjunto de centróides de agrupamento G que apresentou o maior valor para medida de qualidade $\mathcal{H}$.

1: $(\mathbf{V}^A, \mathbf{C}^A) \leftarrow$ **AmostragemEstratificada**$(\mathbf{V}, \mathbf{C})$
2: $\mathbf{V}^A \leftarrow \mathbf{V} - \mathbf{V}^A$, $\mathbf{C}^B \leftarrow \mathbf{C} - \mathbf{C}^A$
3: $G_R \leftarrow$ **MedóidesAleatórios**$(\mathbf{V}^A, k_0)$
4: $G_0 \leftarrow$ **kMeans**$(\mathbf{V}^A, G_R)$
5: $\mathcal{Q} \leftarrow$ **AvaliaAgrupamentos**$(G_0, \mathbf{V}^B, \mathbf{C}^B)$
6: $i \leftarrow 0$
7: **repetir**
8: $\quad i \leftarrow i + 1$
9: $\quad G_{\text{temp}} \leftarrow$ **DivideCentróides**$(G_{i-1}, \mathbf{V}^B, \mathbf{C}^B)$
10: $\quad G_i \leftarrow$ **kMeans**$(\mathbf{V}^A, G_{\text{temp}})$
11: $\quad \mathcal{Q} \leftarrow$ **AvaliaAgrupamentos**$(G_i, \mathbf{V}^B, \mathbf{C}^B)$
12: **até que** $|G_i| < k_{\max}$
13: **retorna** $\{G_i | \mathcal{Q}_i = \max\{\mathcal{Q}_1, \mathcal{Q}_2, \cdots\}\}$

Na linha 3 o procedimento **MedóidesAleatórios** retorna k_0 medóides aleatórios. Os medóides retornados são k_0 vetores de características aleatoriamente escolhidos de $\mathbf{V}^A$ que serão utilizados como os centróides iniciais para execução do algoritmo.

O procedimento **kMeans** na linha 4 retorna a posição dos centróides tomando como entrada os vetores de características $\mathbf{V}^A$ e como centróides iniciais G_R. Na linha 5, o conjunto de centróides retornado, denotado por G_0, é avaliado utilizando o subconjunto $\mathbf{V}^B$ de vetores de características e os respectivos valores de atributo classe. Para tanto, é empregada a função **AvaliaAgrupamentos** que calcula a entropia de distribuição de classe para cada agrupamento por meio da equação 5.4:

$$\mathcal{H}(\theta) = -\sum_{c' \in \mathcal{C}} p(c'|\theta) \log p(c'|\theta) \tag{5.4}$$

onde θ denota o agrupamento, $\mathcal{C}$ denota o conjunto de todos os possíveis valores para o atributo classe e $p(c'|\theta)$ corresponde a freqüência do valor de classe c' para os vetores de características em $\mathbf{V}^B$ que foram associados ao agrupamento θ. A medida de qualidade retornada para o conjunto de centróides é denotada por $\mathcal{Q}$ e corresponde ao inverso multiplicativo da média das entropias de cada agrupamento ponderada pelo número de vetores associados a cada agrupamento. Ou seja:

$$\mathcal{Q} = \left[\frac{1}{\sum_{\theta \in \Theta} |\theta|} \sum_{\theta \in \Theta} |\theta| H(\theta) \right]^{-1} \tag{5.5}$$

Na equação 5.5 Θ corresponde ao conjunto de todos os agrupamentos e $|\theta|$ ao número de elementos associado ao agrupamento θ. Nas linhas 7-12 do algoritmo 5.5 os centróides que apresentam os menores valores de entropia são divididos enquanto o número de agrupamentos resultantes é menor que $k_{\max}$. O procedimento **DivideCentróides** na linha 9 avalia a entropia de cada agrupamento individualmente. Os agrupamentos são então ordenados com respeito ao valor de entropia e a metade dos agrupamentos que apresentarem os maiores valores (ou seja, que são mais impuros) são divididos. O processo de divisão consiste em escolher aleatoriamente dois vetores associados aos centróides para serem os novos centróides. Por fim, na linha 13 o conjunto de centróides que apresentou o maior valor para a medida de qualidade é retornado.

É importante notar que o algoritmo de agrupamento proposto depende da definição dos valores mínimo e máximo do parâmetro k. Empiricamente, constatou-se que utilizar um valor de k_0 menor que o número de classes do conjunto de dados não resulta em uma melhora no desempenho de classificação do método *Concept*. Adicionalmente, valores de $k_{\max}$ maiores que quatro vezes o número de classe tornam o processo de agrupamento lento e não resultam em um aumento significativo do desempenho de classificação. Desta maneira, foi adotado nos experimentos realizados (seção 5.4) o valor de k_0 igual ao número de classes do conjunto de imagens e $k_{\max} = 4k_0$.

5.4 Experimentos

Nesta seção serão descritos os experimentos realizados com o propósito de validar o método *Concept*. A implementação do método *Concept* e do método *IDEA* foi realizada utilizando a linguagem de programação C++. Para mineração de regras de associação foi utilizada para ambos os algoritmos a implementação na linguagem C do algoritmo *Apriori* descrita em [Borgelt 05]. Para os demais classificadores considerados nos experimentos foi utilizada a implementação disponibilizada pelo software de mineração de dados Weka [Hall 09].

Para todos os experimentos foi utilizada validação cruzada com dez partições. Sendo assim, foram realizadas dez repetições para cada experimento, sendo que nove partições são utilizadas para induzir o modelo de classificação e a partição restante é utilizada para validação. Adicionalmente, adotou-se a metodologia de utilizar amostragem estratificada para gerar as partições. Desta maneira, a distribuição de classe em cada uma das partições é aproximadamente igual a do conjunto original.

5.4.1 Classificação de Imagens de Mamografia

Nesta seção são descritos experimentos realizados para a tarefa de classificação de imagens de mamografia utilizando um conjunto de imagens de regiões de interesse (*Regions of Interest - ROIs*) de mamografias disponibilizadas pelo departamento de radiologia da Universidade de Viena (*Department of Radiology of University of Vienna*). O conjunto de imagens foi empregado pelos autores do método *IDEA* em sua avaliação [Ribeiro 09] e por este motivo foi utilizado para realizar uma comparação de desempenho entre o método *IDEA* e o método *Concept*. O conjunto de imagens consiste em 446 *ROIs* de tecidos tumorais obtidos de mamografias que são utilizadas por um sistema de treinamento de estudantes de radiologia[1].

Cada uma das imagens têm associado um valor de *BI-RADS* (*Breast Imaging Reporting Data System*). Os valores de *BI-RADS* compõem uma escala que foi desenvolvida pelo Colégio Americano de

[1] `www.birads.at`

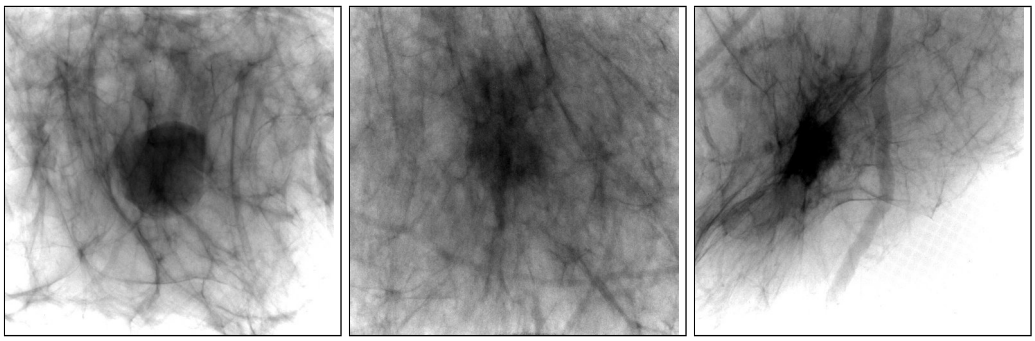
Radiologia (*American College of Radiology*) para padronizar os laudos e os procedimentos de diagnóstico de mamografias. É importante notar que neste conjunto de imagens são encontradas *ROIs* com valor de *BI-RADS* entre três e cinco. A tabela 5.2 mostra a distribuição dos valores de *BI-RADS* para o conjunto de imagens. O significado de cada valor de *BI-RADS* é apresentado na tabela 5.1 e exemplos de *ROIs* para cada um desses valores são apresentadas na figura 5.5.

Tabela 5.1: Níveis de *BI-RADS* e suas descrições.

BI-RADS	**Descrição**
0	Necessidade de exame complementar.
1	Normal.
2	Achados benignos.
3	Achado provavelmente benigno.
4	Achado supeito de malignidade.
5	Achado altamente suspeito de malignidade.

Tabela 5.2: Distribuição de classes (valores de *BI-RADS*) para o conjunto de imagens *ROIs Vienna*.

BI-RADS	**Número de *ROIs***
3	108
4	232
5	106



(a) (b) (c)

Figura 5.5: Exemplos de *ROIs* do conjunto de imagens de mamografias da Universidade de Viena. Notar que os valores de níveis de cinza foram invertidos para facilitar a visualização. (a) *BI-RADS* 3. (b) *BI-RADS* 4. (c) *BI-RADS* 5.

O processo de extração de características para este experimento é o mesmo adotado em [Ribeiro 09] que consistiu em segmentar as imagens e extrair características de textura, forma e cor das regiões evidenciadas. O processo de segmentação foi dividido em duas etapas. Inicialmente, foi realizada segmentação por limiarização para remover pixels com valor menor que 0,14 em uma escala de 0,0 a

1,0 e então aplicado o algoritmo de Otsu [Otsu 79] na imagem resultante para realizar uma segunda limiarização. As características extraídas foram:

Distribuição dos níveis de cinza (8 componentes): Média, desvio padrão, contraste, obliqüidade (*skewness*), curtose e entropia;

Sumarizações da matriz de co-ocorrência de níveis de cinza (4 componentes): Foi extraída a matriz de co-ocorrência de níveis de cinza da região segmentada utilizando ângulo de zero grau e distância entre pixels igual a um. Da matriz resultante foram calculadas as sumarizações de contraste, correlação, energia e homogeneidade. As sumarizações correspondem aos quatro primeiros descritores de Haralick [Haralick 79];

Momentos Geométricos (6 componentes): Correspondem aos momentos geométricos propostos em [Hu 62]. São utilizados para descrever a forma de regiões segmentadas da imagem.

Para o método *IDEA* as características foram organizadas na forma de um vetor de características de 18 componentes. Para o classificador *Concept*, as características foram divididas em três espaços de características:

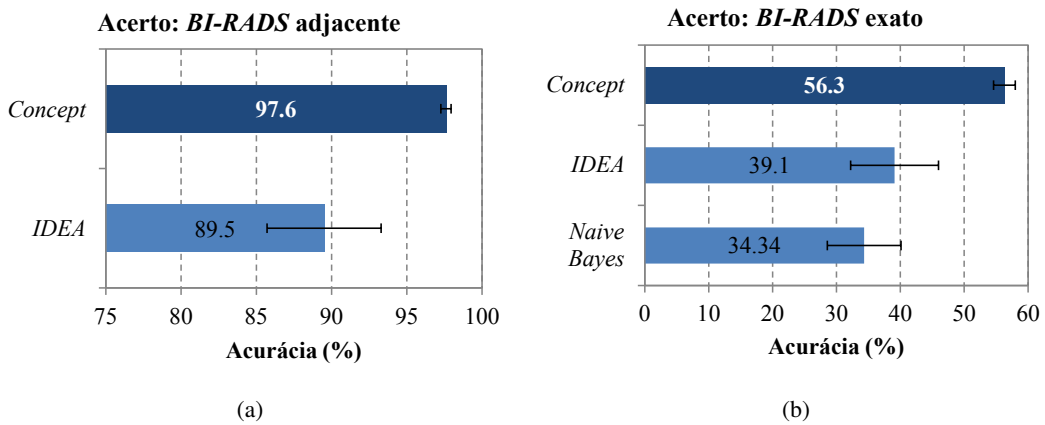
1. distribuição dos níveis de cinza;
2. sumarizações da matriz de co-ocorrência;
3. momentos de geométricos.

Na avaliação feita em [Ribeiro 09] uma predição é considerada correta se o valor de *BI-RADS* difere em uma unidade ou é igual ao valor de classe correto para a imagem. Os resultados obtidos utilizando a metodologia descrita são apresentados no gráfico da figura 5.6(a). Deve-se notar que para esta metodologia um classificador estocástico, que retorna um valor de classe aleatório assumindo distribuição uniforme, terá acurácia esperada de 84,08% para este conjunto de imagens. Adicionalmente, um classificador que sempre classifique as imagens com o valor de *BI-RADS* igual a quatro terá 100% de acertos, o que não é correto.

O gráfico da figura 5.6(b) apresenta os resultados para essa base de imagens, considerando acertos somente se o valor de *BI-RADS* predito é igual ao valor correto para a imagem. Também é incluído para comparação os resultados obtidos com o classificador *Naive Bayes* [John 95]. Assim como para o *IDEA*, as características extraídas foram organizadas como um vetor de características de 18 componentes. Para ambas as abordagens o classificador *Concept* apresentou um desempenho superior considerando o mesmo conjunto de imagens empregado pelos autores do método *IDEA* para avaliá-lo.

5.4.2 Classificação de Doenças Pulmonares

Os experimentos descritos nesta seção tiveram como objetivo comparar o desempenho do método *Concept* e o método *IDEA* para a tarefa de classificação de doenças pulmonares. Para tanto, foi utilizada a base de imagens pulmonares descrita na seção 4.3. São imagens de tomografias do pulmão separadas em *ROIs* de 64×64 pixels. Cada uma das *ROIs* foi classificada por um especialista médico como normal ou um padrão de doença. Os padrões de doença considerados foram: (i) enfisema, (ii) consolidação,



(a) (b)

Figura 5.6: Acurácia de classificação para o conjunto de imagens *ROIs Vienna*. As barras de erro indicam um desvio padrão. (a) Resultados considerando como acerto valores de *BI-RADS* adjacentes ao predito. (b) Resultados considerando como acerto valores de *BI-RADS* iguais ao predito.

(iii) espessamento, (iv) favo de mel e (v) vidro fosco. De cada *ROI* foram extraídas as seguintes características:

Distribuição dos níveis de cinza (5 Componentes): Mediana, desvio padrão, obliqüidade (*skewness*) e contraste dos primeiros e segundos vizinhos. O contraste dos primeiros e segundos vizinhos corresponde à média de diferença dos níveis de cinza dos pixels distantes em uma e duas unidades entre si;

***FFS* (8 componentes):** Método de extração proposto neste projeto de Mestrado e descrito no capítulo 4;

Haralick (140 componentes): Correspondem às sete primeiras sumarizações das matrizes de co-ocorrência de níveis de cinza propostas em [Haralick 79]. A imagem foi requantizada para 16 níveis de cinza e as matrizes foram calculadas para as distâncias 1, 2, 3, 4 e 5 e para as direções de 0°, 45°, 90°e 135°;

Histograma (16 componentes): Corresponde ao histograma de níveis de cinza extraído da imagem após sua requantização para 16 níveis de cinza;

Momentos de Zernike (256 componentes): Os momentos de Zernike [Khotanzad 90] são utilizados para descrever características de forma de uma imagem. Sua principal vantagem é que a etapa de segmentação é dispensada.

Nesta avaliação foram considerados os classificadores associativos *IDEA* e *Concept*. Adicionalmente, para comparação, também são incluídos os resultados dos classificadores *Naive Bayes* e *kNN*. Para o *Concept* as características foram divididas em cinco espaços de características:

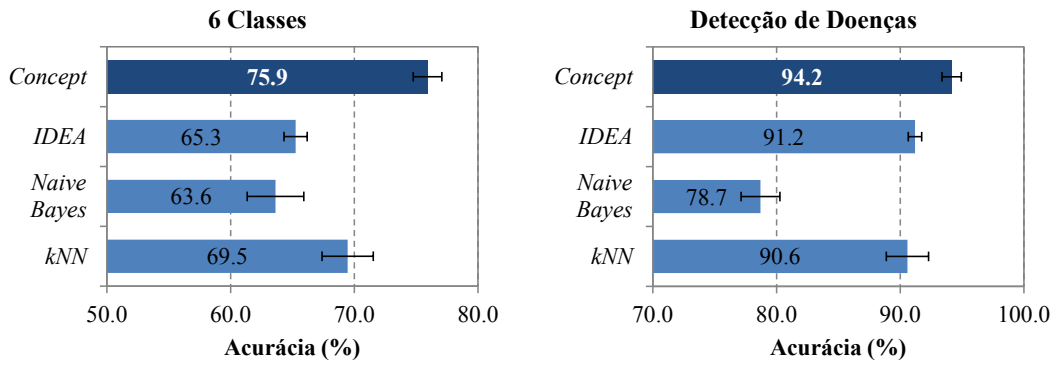
1. distribuição dos níveis de cinza;
2. descritores extraídos pelo *FFS*;
3. descritores de Haralick;

4. histograma;

5. momentos de Zernike.

Para os demais classificadores as características extraídas foram organizadas em um vetor de características.

No primeiro experimento realizado foi avaliado o desempenho para a tarefa de classificação das seis diferentes classes presentes no conjunto de imagens. Os resultados são mostrados na figura 5.7(a). No segundo experimento foram consideradas duas classes: *ROIs* normais e não-normais (*ROIs* com presença de algum padrão de doença pulmonar). Os resultados do segundo experimento são dados na figura 5.7(b).



(a) (b)

Figura 5.7: Acurácia de classificação para o conjunto de imagens *ROIs* do pulmão do Hospital das Clínicas de Ribeirão Preto. As barras de erro indicam um desvio padrão. (a) Resultados considerando seis classes. (b) Resultados considerando duas classes: *ROIs* normais e não-normais.

Para ambos os experimentos o método *Concept* obteve uma acurácia de classificação significativamente maior que os demais classificadores. É importante notar que no segundo experimento a distribuição de classes não é balanceada. Analisando a distribuição das 6 classes do conjunto (dada na tabela 4.1 no capítulo 4) observa-se que existem 590 *ROIs* normais contra 2.668 *ROIs* não-normais. Desta maneira, o *Concept* se mostrou superior aos demais classificadores mesmo para o conjunto de dados não balanceado.

A tabela 5.3 mostra além dos valores de acurácia de classificação, a precisão e especificidade (taxa de verdadeiros negativos - TVN) obtida com cada classificador para as seis classes. O desvio padrão é indicado entre parênteses e os valores que são significativamente maiores que os demais, considerando um teste t-Student com $p = 0,01$, são marcados com $*$. Os resultados obtidos considerando duas classes são dados na tabela 5.4. Com exceção da especificidade para a classificação com duas classes, na qual o classificador *Naive Bayes* apresentou um valor maior, o *Concept* demonstrou desempenho superior com respeito às medidas de acurácia, precisão e especificidade.

Tabela 5.3: Desempenho de classificação para o conjunto de imagens de *ROIs* do pulmão do Hospital das Clínicas de Ribeirão Preto considerando seis classes. Os valores entre parênteses indicam desvio padrão. Os valores marcados com $*$ indicam valores estatisticamente superiores aos demais com $p = 0,01$.

	6 Classes		
	Acurácia (%)	**Precisão (%)**	**Especificidade (%)**
kNN	69,5 (2,1)	70,0 (2,3)	93,7 (0,4)
Naive Bayes	63,6 (2,3)	63,0 (2,1)	92,4 (0,5)
IDEA	65,3 (0,9)	64,8 (0,9)	93,1 (0,2)
Concept	$*$ 75,9 (1,2)	$*$ 76,0 (1,1)	$*$ 95,1 (0,2)

Tabela 5.4: Desempenho de classificação para o conjunto de imagens de *ROIs* do pulmão do Hospital das Clínicas de Ribeirão Preto considerando duas classes. Os valores entre parênteses indicam desvio padrão. Os valores marcados com $*$ indicam valores estatisticamente superiores aos demais com $p = 0,01$.

	2 Classes		
	Acurácia (%)	**Precisão (%)**	**Especificidade (%)**
kNN	90,6 (1,7)	91,3 (1,3)	83,7 (3.1)
Naive Bayes	78,7 (1,6)	88,2 (0,9)	$*$ 88,6 (2,4)
IDEA	91,2 (0,5)	91,5 (0,5)	81,8 (3,1)
Concept	$*$ 94,2 (0,8)	$*$ 94,0 (0,8)	81,0 (3,5)

5.5 Conclusões

Nesta seção foi apresentado o novo método de classificação de imagens desenvolvido durante este trabalho e denominado *Concept*. O *Concept* é um classificador associativo, ou seja, que utiliza técnicas de mineração de regras de associação para prever a classe de uma imagem.

A principal contribuição do método refere-se ao algoritmo de obtenção de representação das imagens. Os classificadores associativos existentes na literatura empregam técnicas de discretização para transformar os vetores de características em *itemsets*. Já o *Concept* utiliza um novo algoritmo, o *MFS-Map*, que realiza agrupamento de dados em diferentes espaços de características para obter os *itemsets* que irão representar as imagens. A principal vantagem do *MFS-Map* está no fato de que os itens obtidos representam regiões dos espaços de características. Em tais regiões as imagens são visualmente similares e, desta maneira, os itens obtidos pelo *MFS-Map* carregam informações semânticas valiosas para o processo de classificação.

O *Concept* foi avaliado para diferentes tarefas envolvendo imagens médicas, tais como a predição de doenças pulmonares e classificação de lesões presentes em imagens de mamografia. Em todos os experimentos realizados o desempenho obtido pelo *Concept* se mostrou superior ao do método *IDEA*, que foi utilizado como base para o desenvolvimento deste projeto de Mestrado e que também utiliza mineração de regras de associação para realizar a classificação de imagens. Os resultados obtidos indicam que a abordagem de representações adotada no classificador *Concept*, que dispensa a discretização de atributos, é promissora para a classificação associativa de imagens.

Parte III

Conclusões

Capítulo 6

Conclusões Gerais e Linhas de Futuras Pesquisas

O volume de dados armazenados em sistemas computacionais da área médica, que incluem laudos compostos de textos e diversas modalidades de imagens médicas, tem uma tendência de crescimento exponencial. Esse grande volume de dados é uma valiosa fonte de conhecimentos que pode ser aplicada para o auxílio ao diagnóstico médico e para o ensino da medicina. No entanto, em virtude da complexidade da análise e tratamento destes dados é crucial o desenvolvimento de técnicas computacionais para extrair todo o seu potencial. De fato, pesquisas mostram que o diagnóstico auxiliado por computador (*Computer Aided Diagnosis - CAD*) pode melhorar significativamente o desempenho de radiologistas em tarefas como a detecção de anomalias em mamografias [Doi 07].

A configuração típica de um sistema *CAD* apoiado por mineração de imagens consiste na extração das informações visuais relevantes de uma imagem na forma de vetores de características. Esses vetores são utilizados como entrada para um classificador ou algum outro método de mineração de dados. No entanto, devido ao problema conhecido como lacuna semântica [Deserno 09], que corresponde à diferença existente entre a percepção da imagem pelo especialista médico e as características automaticamente extraídas, um aspecto desafiador da tarefa de mineração de imagens médicas consiste em obter características que descrevam adequadamente as informações visuais de imagens médicas.

Assim, este trabalho teve como primeiro foco o desenvolvimento de algoritmos para extração de características de formas voltadas para o domínio de imagens médicas. A hipótese considerada foi que a forma de objetos presentes na imagem é um atributo visual que aproxima a semântica esperada pelo especialista médico. O segundo foco consistiu no desenvolvimento de técnicas de mineração voltadas para a análise das informações contidas nas características extraídas das imagens.

6.1 Principais Contribuições

Foram duas as principais contribuições deste projeto de Mestrado. A primeira contribuição consistiu na técnica de extração de características denominada *FFS* (*Fast Fractal Stack*). O *FFS* emprega análise fractal para medir a complexidade dos contornos e objetos presentes em imagens, retornando um vetor de características compacto e com alto poder descritivo. Foram realizados experimentos para a tarefa de classificação de doenças em imagens de tomografia do pulmão. Os resultados obtidos demonstraram a eficácia do *FFS* em classificar pulmões saudáveis com respeito a cinco diferentes padrões de doenças pulmonares. Adicionalmente, o *FFS* apresenta um algoritmo de extração eficiente, com custo linear com respeito ao número de pixels da imagem. Os resultados obtidos com o extrator foram publicados no *ACM Workshop on Medical Multimedia Analysis and Retrieval* (*MMAR 2011*) que ocorreu juntamente com a conferência *ACM Multimedia 2011* [Costa 11].

A segunda contribuição deste trabalho foi o classificador de imagens *Concept*. Trata-se de um classificador associativo, ou seja, que emprega mineração de regras de associação para predizer a classe de uma imagem. O aspecto inovador do *Concept* refere-se ao algoritmo de obtenção de representação de imagens. Os classificadores associativos existentes na literatura empregam técnicas de discretização para transformar vetores de características extraídos das imagens em *itemsets*. Já o *Concept* utiliza um novo algoritmo desenvolvido neste trabalho, o *MFS-Map*, que realiza agrupamento de dados em diferentes espaços de características para obter os *itemsets* que irão representar as imagens.

O *Concept* foi avaliado para diferentes tarefas envolvendo imagens médicas, tais como a predição de doenças pulmonares e classificação de lesões presentes em imagens de mamografia. Em todos os experimentos realizados o *Concept* apresentou desempenho superior ao método de classificação associativa *IDEA* (seção 2.3) que emprega discretização de atributos para obter os *itemsets* que representam as imagens no processo de classificação. Os resultados obtidos indicam que os itens obtidos pelo algoritmo *MFS-Map* carregam informações que são mais vantajosas para o processo de classificação.

6.2 Linhas de Futuras Pesquisas

Os trabalhos realizados ao longo deste projeto de Mestrado possibilitam o desenvolvimento de extensões das técnicas propostas. As principais linhas de futuras pesquisas são listadas a seguir:

Avaliar *FFS* para Novos Domínios de Aplicação: O algoritmo *FFS* foi avaliado neste trabalho para a classificação de imagens médicas. No entanto, o método apresenta características desejáveis que poderiam ser úteis em outros domínios de aplicação. Por exemplo, a eficiência do algoritmo de extração, que tem custo linear com relação ao número de pixels da imagem, poderia ser útil para sistemas de recuperação de imagens por conteúdo interativos.

Investigar Técnicas de Segmentação para o Extrator *FFS*: O algoritmo *FFS* emprega a técnica de decomposição em pilha de imagens binárias para particionar a imagem a ser descrita em um conjunto de imagens binárias. Para tanto, a imagem é limiarizada utilizando sucessivos valores de limiar. Assim, investigar outras técnicas de segmentação que poderiam ser empregadas para gerar a pilha de imagens binárias poderia levar a um aumento da eficácia do *FFS*.

Avaliar Algoritmos de Agrupamento Eficientes para o *MFS-Map*: O *MFS-Map* emprega o algoritmo de análise de agrupamentos descrito na seção 5.3 que é baseado na extensão do algoritmo *k-Means* proposta em [Pelleg 00]. Para bases de imagens de alta cardinalidade, ou seja, que possuem um grande número de imagens, o custo computacional de execução do algoritmo de agrupamento pode ser inaceitável. Desta maneira, uma linha de pesquisa futura consiste em avaliar algoritmos de agrupamento mais eficientes para o algoritmo *MFS-Map*.

Referências Bibliográficas

[Agrawal 93] R. Agrawal, T. Imielinki & A. Swami. *Mining association rules between sets of items in large databases*. ACM SIGMOD Record, pages 207–216, 1993.

[Agrawal 94] R. Agrawal & R. Srikant. *Fast algorithms for mining association rules*. In 20th International Conference on Very Large Data Bases, VLDB, pages 487–499, Santiago de Chile, Chile, 1994. Morgan Kaufmann.

[Antonie 04] M.L. Antonie & O.R. Zaïane. *An associative classifier based on positive and negative rules*. In Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pages 64–69, Paris, France, 2004. ACM.

[Backes 10] A. Backes & O. Bruno. *Shape Skeleton Classification Using Graph and Multi-scale Fractal Dimension*. Image and Signal Processing, no. i, pages 448–455, 2010.

[Baralis 02] E. Baralis & P. Garza. *A lazy approach to pruning classification rules*. In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), pages 35–42, IEEE Computer Society, 2002.

[Becker 10] H. Becker, M. Naaman & L. Gravano. *Learning similarity metrics for event identification in social media*. In Proceedings of the third ACM international conference on Web search and data mining, pages 291–300. ACM, 2010.

[Beucher 79] S. Beucher & C. Lantuejoul. *Use of Watersheds in Contour Detection*. In International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation, Rennes, France., September 1979.

[Bick 99] U. Bick & H. Lenzen. *PACS: the silent revolution*. European Radiology, vol. 1160, pages 1152–1160, 1999.

[Blum 67] H. Blum. *A transformation for extracting new descriptors of shape*. Models for the perception of speech and visual form, vol. 19, no. 5, pages 362–380, 1967.

[Borgelt 05] C. Borgelt. *An implementation of the FP-growth algorithm*. In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, pages 1—-5, Chicago, Illinois, 2005. ACM Press.

[Bruno 08] O. M. Bruno, R. O. Plotze, M. Falvo & M. Castro. *Fractal dimension applied to plant identification*. Information Sciences, vol. 178, no. 12, pages 2722–2733, June 2008.

[Bugatti 09] Pedro H. Bugatti, M. P. Silva, A. J. M. Traina, C. Traina Jr. & P. M. A. Marques. *Content-based retrieval of medical images: From context to perception*. In 22nd IEEE International Symposium on Computer-Based Medical Systems, pages 1–8, Albuquerque, USA, August 2009. IEEE.

[Burl 99] M. C. Burl, C. Fowlkes & J. Roden. *Mining for image content*. Rapport technique, 1999.

[Canny 86] J. Canny. *A computational approach to edge detection*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 8, no. 6, pages 679–698, November 1986.

[Chan 90] H. P. Chan, K. Doi, C. J. Vybrony, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu & H. Macmahon. *Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms: The Potential of Computer-Aided Diagnosis*. Investigative Radiology, vol. 25, no. 10, pages 1102–1110, 1990.

[Chen 95] Y. Q. Chen & M. S. Nixon. *Statistical geometrical features for texture classification*. Pattern Recognition, vol. 28, no. 4, pages 537–552, 1995.

[Comer 94] M. L. Comer & E. J. Delp. *Parameter estimation and segmentation of noisy or textured images using the EM algorithm and MPM estimation*. Image Processing, 1994. Proceedings., vol. 2, no. 31 7, pages 650–654, 1994.

[Costa 99] L. F. Costa & L. F. Estrozi. *Multiresolution shape representation without border shifting*. Electronics Letters, vol. 35, no. 21, page 1829, 1999.

[Costa 01] L. F. Costa & R. M. Cesar Jr. Shape Analysis and Classification: Theory and Pratice. CRC Press, Boca Raton, Florida, USA, 2001.

[Costa 11] A. F. Costa, J. Tekli & A. J. M. Traina. *Fast Fractal Stack: Fractal Analysis of Computed Tomography Scans of the Lung*. In MMAR'11:International ACM Workshop on Medical Multimedia Analysis and Retrieval, pages 1–6, Scottsdale, AZ, USA, 2011. ACM.

[Datta 08] Ritendra Datta, Dhiraj Joshi, Jia Li & James Z. Wang. *Image Retrieval: Ideas, Influences, and Trends of the New Age*. ACM Computing Surveys, vol. 40, no. 2, pages 1–60, 2008.

[Depeursinge 08] A. Depeursinge, J. Iavindrasana, A. Hidki, G. Cohen, A. Geissbuhler, A. Platon, P.A. Poletti & H. Muller. *A classification framework for lung tissue categorization*. In SPIE Medical Imaging, volume 41, pages 69190C–69190C–12. SPIE, 2008.

[Depeursinge 10] A. Depeursinge, D. Racoceanu, J. Iavindrasana, G. Cohen, A. Platon, P. A. Poletti & H. Müller. *Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography.* Artificial intelligence in medicine, vol. 50, pages 13–21, May 2010.

[Deriche 93] R. Deriche & G. Giraudon. *A computational approach for corner and vertex detection*. International Journal of Computer Vision, vol. 10, no. 2, pages 101–124, 1993.

[Deserno 09] T. M. Deserno, S. Antani & R. Long. *Ontology of gaps in content-based image retrieval*. Journal of digital imaging, vol. 22, no. 2, pages 202–15, April 2009.

[Doi 07] K. Doi. *Computer-aided diagnosis in medical imaging: historical review, current status and future potential.* Computerized medical imaging and graphics, vol. 31, no. 4-5, pages 198–211, 2007.

[Falcão 02] A. X. Falcão, L. F. Costa & B. S. Cunha. *Multiscale skeletons by image foresting transform and its application to neuromorphometry*. Pattern Recognition, vol. 35, pages 1571–1582, 2002.

[Fischer 08] B. Fischer, T. M. Deserno, B. Ott & R. W. Günther. *Integration of a research CBIR system with RIS and PACS for radiological routine*. Proceedings of SPIE, pages 691914–691914–10, 2008.

[Giger 08] M. L. Giger, H. P. Chan & J. Boone. *Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM*. Medical Physics, vol. 35, no. 12, page 5799, 2008.

[Granlund 72] G. H. Granlund. *Fourier preprocessing for hand print character recognition*. Computers, IEEE Transactions on, pages 195–201, 1972.

[Hall 00] M. A. Hall. *Correlation-based feature selection for discrete and numeric class machine learning*. In 17th International Conference on Machine Learning, pages 359–366, Stanford, CA, 2000.

[Hall 09] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten. *The WEKA data mining software: an update*. ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, page 10, November 2009.

[Han 00] J. Han, J. Pei & Y. Yin. *Mining frequent patterns without candidate generation*. ACM SIGMOD Record, vol. 29, no. 2, pages 1–12, 2000.

[Han 03] Y. Han. *CPAR: Classification based on predictive association rules*. In Proceedings of SIAM international conference on data mining, pages 331–335, 2003.

[Han 05] J. Han & M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[Haralick 73] R. M. Haralick, K. Shanmugam & I. H. Dinstein. *Textural Features for Image Classification*. IEEE Transactions on Systems, Man and Cybernetics, vol. 3, no. 6, pages 610–621, 1973.

[Haralick 79] R. M. Haralick. *Statistical and structural approaches to texture*. In Proceedings of the IEEE, volume 67, pages 786–804, 1979.

[Heath 00a] M. D. Heath & K. W. Bowyer. *Computer Aided Detection for Screening Mammography*. In Handbook of Image and Video Processing, chapitre 10.4, pages 805–820. Academic Press, Orlando, FL, USA, 1st edition, 2000.

[Heath 00b] M.D. Heath, K. W. Bowyer, D. Kopans, R. Moore & W. P. Kegelmeyer. *The Digital Database for Screening Mammography*. In M. J. Yaffe, editeur, Fifth International Workshop on Digital Mammography, pages 212–218, Toronto, Canada, 2000. Medical Physics Publishing.

[Hsu 02] W. Hsu, M. L. Lee & J. Zhang. *Image mining: trends and developments*. Journal of Intelligent Information Systems, pages 7–23, 2002.

[Hu 62] M. K. Hu. *Visual pattern recognition by moment invariants*. Information Theory, IRE Transactions on, vol. 8, no. 2, pages 179–187, 1962.

[Huber 10] M. B. Huber, M. Nagarajan, G. Leinsinger, L. A. Ray & A. Wismuller. *Classification of interstitial lung disease patterns with topological texture features*. In Proceedings SPIE Medical Imaging 2010, volume 7624, pages 2–9, 2010.

[Inamura 95] K. Inamura & T. Takahashi. *Storage and presentation of images.* International journal of bio-medical computing, vol. 39, no. 1, pages 157–62, April 1995.

[Jain 10] Anil K. Jain. *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, vol. 31, no. 8, pages 651–666, June 2010.

[John 95] G.H. John & P. Langley. *Estimating continuous distributions in Bayesian classifiers*. In Proceedings of the eleventh conference on uncertainty in artificial intelligence, volume 1, pages 338–345. Citeseer, 1995.

[Kass 88] M. Kass, A. Witkin & D. Terzopoulos. *Snakes: Active contour models*. International journal of computer vision, vol. 1, no. 4, pages 321–331, 1988.

[Khotanzad 90] A. Khotanzad & Y. H. Hong. *Invariant image recognition by Zernike moments*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pages 489–497, May 1990.

[Khutlang 10] R. Khutlang, S. Krishnan, R. Dendere, A. Whitelaw, K. Veropoulos, G. Learmonth & T. S. Douglas. *Images of ZN-Stained Sputum Smears*. IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pages 949–957, 2010.

[Kinoshita 07] S. K. Kinoshita, P. M. A. Marques, R. R. Pereira, J. A. H. Rodrigues & R. M. Rangayyan. *Content-based retrieval of mammograms using visual features related to breast density patterns.* Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology, vol. 20, no. 2, pages 172–90, June 2007.

[Kiranyaz 10] S. Kiranyaz & M. Birinci. *Perceptual color descriptor based on spatial distribution: A top-down approach*. Image and Vision Computing, vol. 28, no. 8, pages 1309–1326, 2010.

[Kriegel 09] H. P. Kriegel, P. Kröger & A. Zimek. *Clustering high-dimensional data*. ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 1, pages 1–58, March 2009.

[Liu 98] B. Liu, W. Hsu & Y. Ma. *Integrating classification and association rule mining*. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), pages 80–86, New York, NY, USA, 1998. AAAI Press.

[Liu 07] Y. Liu, D. Zhang, G. Lu & W. Ma. *A survey of content-based image retrieval with high-level semantics*. Pattern Recognition, vol. 40, no. 1, pages 262–282, January 2007.

[Loizou 11] C. P. Loizou, V. Murray, M. S. Pattichis, I. Seimenis, M. Pantziaris & C. S. Pattichis. *Multiscale Amplitude-Modulation Frequency-Modulation (AM-FM) Texture Analysis of Multiple Sclerosis in Brain MRI Images.* IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 1, pages 119–29, January 2011.

[Ma 99] W. Y. Ma & B. S. Manjunath. *NeTra: A toolbox for navigating large image databases*. Multimedia Systems, vol. 7, no. 3, pages 184–198, 1999.

[Mandelbrot 83] B. B. Mandelbrot. The Fractal Geometry of Nature. Times Books, 1st editio edition, 1983.

[Marques 09] P. M. A. Marques & S. C. Salomão. *PACS: Sistemas de Arquivamento e Distribuição de Imagens*. Revista Brasileira de Física, vol. 3, no. 1, pages 131–139, 2009.

[Marr 80] D. Marr & E. Hildreth. *Theory of Edge Detection*. Proceedings of the Royal Society of London. Series B, Biological Sciences, vol. 207, no. 1167, pages 187–217, 1980.

[Naqa 05] I. E. Naqa & Y. Yang. *Techniques in the detection of microcalcification clusters in digital mammograms*. In Medical Imaging Systems: Technology and Applications, pages 15–36. World Scientific, Singapore, 2005.

[Otsu 79] N. Otsu. *A threshold selection method from gray-level histograms*. IEEE Transactions on Systems, Man and Cybernetics, vol. 9, no. 1, pages 62–66, 1979.

[Pang-Ning 05] T. Pang-Ning, M. Steinbach & K. Vipin. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, first edit edition, 2005.

[Pelleg 00] D. Pelleg & A. Moore. *X-means: Extending k-means with efficient estimation of the number of clusters*. In Proceedings of the Seventeenth International Conference on Machine Learning, volume 1, pages 727–734. San Francisco, 2000.

[Platt 99] J.C. Platt. *Fast training of support vector machines using sequential minimal optimization*. In Advances in Kernel Methods, chapitre 12, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[Prati 08] R. C. Prati, G. Batista & M. C. Monard. *Curvas ROC para a avaliação de classificadores*. Revista IEEE América Latina, vol. 6, no. 2, pages 215–222, 2008.

[Rangayyan 04] R. M. Rangayyan. Biomedical Image Analysis. CRC Press, Boca Raton, Florida, USA, 2004.

[Rangayyan 07] R. M. Rangayyan & T. M. Nguyen. *Fractal analysis of contours of breast masses in mammograms*. Journal of Digital Imaging, vol. 20, no. 3, pages 223–37, September 2007.

[Ribeiro 08] M. X. Ribeiro, A. J. M. Traina, C. Traina Jr., N. A. Rosa & P. M. A. Marques. *How to Improve Medical Image Diagnosis through Association Rules: The IDEA Method*. In 21th IEEE International Symposium on Computer-Based Medical Systems, pages 266–271, Jyväskylä, Finland, June 2008. IEEE.

[Ribeiro 09] M. X. Ribeiro, Pedro H. Bugatti, C. Traina Jr., P. M. A. Marques, N. A. Rosa & A. J. M. Traina. *Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques*. Data & Knowledge Engineering, vol. 68, no. 12, pages 1370–1382, December 2009.

[Rui 07] X. Rui, M. Li, Z. Li, W. Y. Ma & N. Yu. *Bipartite graph reinforcement model for web image annotation*. Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07, page 585, 2007.

[Schroeder 92] M. Schroeder. Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise. W. H. Freeman, New York, NY, USA, 1992.

[Shen 94] L. Shen, R. M. Rangayyan & J. E. L. Desautels. *Calcifications, Detection And Classification Of Mammographic*. In K. W. Bowyer & S. Astley, editeurs, State of the Art in Digital Mammographic Image Analysis, pages 198–212. World Scientific Publishing Co., Inc., 1994.

[Silva 09] M. P. Silva, A. J. M. Traina, P. M. A. Marques, J. C. Felipe & C. Traina Jr. *Including the perceptual parameter to tune the retrieval ability of pulmonary CBIR systems*. In 22nd IEEE International Symposium on Computer-Based Medical Systems, pages 1–8, Albuquerque, USA, August 2009. IEEE.

[Sluimer 06] Ingrid Sluimer, Arnold Schilham, Mathias Prokop & Bram van Ginneken. *Computer analysis of computed tomography scans of the lung: a survey*. IEEE Transactions on Medical Imaging, vol. 25, no. 4, pages 385–405, April 2006.

[Thabtah 07] Fadi Thabtah. *A review of associative classification mining*. The Knowledge Engineering Review, vol. 22, no. 01, page 37, May 2007.

[Timm 10] F. Timm & T. Martinetz. *Statistical Fourier Descriptors for Defect Image Classification*. In International Conference on Pattern Recognition, volume 1, pages 4198–4201, Istanbul, Turkey, 2010. IEEE.

[Torres 04] R. S. Torres, A. X. Falcão & L. F. Costa. *A graph-based approach for multiscale shape analysis*. Pattern Recognition, vol. 37, no. 6, pages 1163–1174, June 2004.

[Torres 05] R. S. Torres, A. X. Falcão, M.A. Goncalves, B Zhang, W Fan, E.A. Fox & P. Calado. *A new framework to combine descriptors for content-based image retrieval*. portal.acm.org, vol. pages, pages 335–336, 2005.

[Torres 06] R. S. Torres & A. X. Falcão. *Content-based image retrieval: Theory and applications*, 2006.

[Traina Jr. 00] C. Traina Jr., A. J. M. Traina, Leejay Wu & C. Faloutsos. *Fast feature selection using fractal dimension*. In Brazilian Symposium on Databases (SBBD), pages 158–171, João Pessoa, Brazil, 2000.

[Uchiyama 03] Y. Uchiyama, S. Katsuragawa, H. Abe, J. Shiraishi, F. Li, Q. Li, C. T. Zhang, K. Suzuki & K. Doi. *Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography*. Medical Physics, vol. 30, no. 9, page 2440, 2003.

[Unay 11] D. Unay, O. Soldea, S. Ozogur-Akyuz, M. Cetin & A. Ercil. *Automated X-Ray Image Annotation*. Multilingual Information Access Evaluation II. Multimedia Experiments, pages 247–254, 2011.

[van De Wetering 09] R. van De Wetering & R. Batenburg. *A PACS maturity model: a systematic meta-analytic review on maturation and evolvability of PACS in the hospital enterprise.* International journal of medical informatics, vol. 78, no. 2, pages 127–40, February 2009.

[Wang 01] J. Z. Wang, J. Li & G. Wiederhold. *SIMPLIcity: Semantics-sensitive integrated matching for picture libraries*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 947–963, 2001.

[Wang 04] X. Wang, M. R. Smith & R. M. Rangayyan. *Mammographic information analysis through association-rule mining*. In Canadian Conference on Electrical and Computer Engineering 2004, pages 1495–1498, Niagara Falls, Canada, 2004. IEEE.

[Wang 09] G. Wang, D. Hoiem & D. Forsyth. *Building text features for object image classification*. In 19th International Conference on Pattern Recognition, pages 1367–1374, Tampa, Florida, USA, June 2009. Citeseer.

[Xu 00] C. Xu, D. Pham & J. Prince. *Image segmentation using deformable models*. In Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis, pages 175–272. SPIE Publications, 2000.

[Zhang 04] D. Zhang. *Review of shape representation and description techniques*. Pattern Recognition, vol. 37, no. 1, pages 1–19, 2004.

[Zhang 08] Hui Zhang, Jason E. Fritts & Sally a. Goldman. *Image segmentation evaluation: A survey of unsupervised methods*. Computer Vision and Image Understanding, vol. 110, no. 2, pages 260–280, May 2008.

[Zhuang 07] Q. Zhuang, J. Feng & H. Bao. *Measuring Semantic Gap: An Information Quantity Perspective*. In 5th IEEE International Conference on Industrial Informatics, volume 100101, pages 669–674, Vienna, 2007. IEEE.