Seleção de características por meio de algoritmos genéticos para aprimoramento de rankings e de modelos de classificação

Sérgio Francisco da Silva

Seleção de características por meio de algoritmos genéticos para aprimoramento de rankings e de modelos de classificação

Sérgio Francisco da Silva

Orientadora: ***Profa. Dra. Agma Juci Machado Traina***
Co-orientador: ***Prof. Dr. João do Espirito Santo Batista Neto***

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

USP – São Carlos
Junho de 2011

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

S586s

Silva, Sérgio Francisco da
Seleção de características por meio de algoritmos genéticos para aprimoramento de rankings e de modelos de classificação / Sérgio Francisco da Silva; orientadora Agma Juci Machado Traina -- São Carlos, 2011.
97 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

1. Seleção de características. 2. Consultas por similaridade. 3. Algoritmos genéticos. 4. Classificação. 5. Imagens médicas. I. Traina, Agma Juci Machado, orient. II. Título.

Este documento foi preparado com o formatador de textos LaTeX 2_ε. O estilo utilizado no documento foi desenvolvido pelo Prof. Dr. Ronaldo Cristiano Prati. A bibliografia foi gerada automaticamente pelo BibTeX, utilizando o estilo *Apalike*. Algumas palavras utilizadas neste documento não foram traduzidas da língua inglesa para a portuguesa por serem amplamente conhecidas e difundidas na comunidade acadêmica.

"*Everything should be made as simple as possible, but not simpler.*"

Albert Einstein

Agradecimentos

Primeiramente a Deus, por estar comigo em todos os momentos da minha vida;

À minha orientadora, Prof^a. Dr^a. Agma J. M. Traina, pela sua sensibilidade e competência em orientar e motivar;

Ao meu co-orientador, Prof. Dr. João Batista Neto, pelos ensinamentos, presteza e atenção;

Aos meus familiares, pelo carinho, compreensão e apoio;

À Aline Kristina pelo amor, carinho e paciência;

Aos meus amigos do Grupo de Bases de Dados e Imagens (GBDI), pelos momentos de estudo e lazer;

Em especial aos meus colegas e amigos Marcos Aurélio, Letrícia e Monica, pela revisão deste texto;

À FAPESP, à CAPES e ao CNPq, pelo apoio financeiro.

Resumo

Sistemas de recuperação de imagens por conteúdo (*Content-based image retrieval* – CBIR) e de classificação dependem fortemente de vetores de características que são extraídos das imagens considerando critérios visuais específicos. É comum que o tamanho dos vetores de características seja da ordem de centenas de elementos. Conforme se aumenta o tamanho (dimensionalidade) do vetor de características, também se aumentam os graus de irrelevâncias e redundâncias, levando ao problema da "maldição da dimensionalidade". Desse modo, a seleção das características relevantes é um passo primordial para o bom funcionamento de sistemas CBIR e de classificação.

Nesta tese são apresentados novos métodos de seleção de características baseados em algoritmos genéticos (do inglês *genetic algorithms* - GA), visando o aprimoramento de consultas por similaridade e modelos de classificação. A família Fc ("*Fitness coach*") de funções de avaliação proposta vale-se de funções de avaliação de *ranking*, para desenvolver uma nova abordagem de seleção de características baseada em GA que visa aprimorar a acurácia de sistemas CBIR. A habilidade de busca de GA considerando os critérios de avaliação propostos (família Fc) trouxe uma melhora de precisão de consultas por similaridade de até 22% quando comparado com métodos *wrapper* tradicionais para seleção de características baseados em *decision-trees* (C4.5), *naive bayes*, *support vector machine*, 1-*nearest neighbor* e mineração de regras de associação.

Outras contribuições desta tese são dois métodos de seleção de características baseados em filtragem, com aplicações em classificação de imagens, que utilizam o cálculo supervisionado da estatística de silhueta simplificada como função de avaliação: o *silhouette-based greedy search* (SiGS) e o *silhouette-based genetic algorithm search* (SiGAS). Os métodos propostos superaram os métodos concorrentes na literatura (CFS, FCBF, ReliefF, entre outros). É importante também ressaltar que o ganho em acurácia obtido pela família Fc, e pelos métodos SiGS e SiGAS propostos proporcionam também um decréscimo significativo no tamanho do vetor de características, de até 90%.

Palavras-chave: Seleção de características; imagens médicas; consultas por similaridade; classificação; algoritmos genéticos.

Abstract

Content-based image retrieval (CBIR) and classification systems rely on feature vectors extracted from images considering specific visual criteria. It is common that the size of a feature vector is of the order of hundreds of elements. When the size (dimensionality) of the feature vector is increased, a higher degree of redundancy and irrelevancy can be observed, leading to the "curse of dimensionality" problem. Thus, the selection of relevant features is a key aspect in a CBIR or classification system.

This thesis presents new methods based on genetic algorithms (GA) to perform feature selection. The Fc ("Fitness coach") family of fitness functions proposed takes advantage of single valued ranking evaluation functions, in order to develop a new method of genetic feature selection tailored to improve the accuracy of CBIR systems. The ability of the genetic algorithms to boost feature selection by employing evaluation criteria (fitness functions) improves up to 22% the precision of the query answers in the analyzed databases when compared to traditional wrapper feature selection methods based on decision-tree (C4.5), naive bayes, support vector machine, 1-nearest neighbor and association rule mining.

Other contributions of this thesis are two filter-based feature selection algorithms for classification purposes, which calculate the simplified silhouette statistic as evaluation function: the silhouette-based greedy search (SiGS) and the silhouette-based genetic algorithm search (SiGAS). The proposed algorithms overcome the state-of-the-art ones (CFS, FCBF and ReliefF, among others). It is important to stress that the gain in accuracy of the proposed methods family Fc, SiGS and SIGAS is allied to a significant decrease in the feature vector size, what can reach up to 90%.

Keywords: Feature selection; medical images; similarity search; classification; genetic algorithms.

Sumário

Lista de Abreviaturas

1NN	*1-Nearest Neighbor*
AM	Aprendizagem de Máquina
CAD	*Computer-Aided Diagnosis*
CBIR	*Content-Based Image Retrieval*
CBR	*Content-Based Retrieval*
CFS	*Correlation-based Feature Selection*
CV	Coeficiente de variação
DICOM	*Digital Imaging and Communication in Medicine*
Fc	*"Fitness coach" function* (função de atribuição de aptidão)
FR-Precision	Função de aptidão derivada da medida *R-Precision*
FCBF	*Fast Correlation Based-Filter*
GA	*Genetic Algorithm*
GA-1NN	*Genetic Algorithm-based feature selection minimizing the 1-Nearest Neighbor classification error*

GA-C4.5	*Genetic Algorithm-based feature selection minimizing the C4.5 classification error*
GA-FcA	*Genetic Algorithm-based feature selection minimizing the FcA criterion*
GA-FcB	*Genetic Algorithm-based feature selection minimizing the FcB criterion*
GA-FcC	*Genetic Algorithm-based feature selection minimizing the FcC criterion*
GA-FR-Precision	*Genetic algorithm-based feature selection minimizing the FR-Precision criterion*
GA-NB	*Genetic Algorithm-based feature selection minimizing the Naive Bayes classification error*
GA-SVM	*Genetic Algorithm-based feature selection minimizing the Support Vector Machine classification error*
GBDI	Grupo de Bases de Dados e Imagens
GS	*Greedy Search*
kNN	*k-Nearest Neighbor*
kNNQ	*k-Nearest Neighbor Query*
kNNGAS	*K-Nearest Neighbor-based Genetic Algorithm Search*
LSD	*Least Significant Difference*
mRMR	*minimal Relevance Maximal Redundance*
MS	*Multistart Search*

MS-FcA	*Multistart Search-based feature selection minimizing the FcA criterion*
NB	*Naive Bayes*
NP	Não polinomial (problema que não pode ser resolvido em tempo polinomial)
PACS	*Picture Archiving and Communication System*
P&R	Precisão e Revocação
PMX	*Partially Matched Crossover*
RBFs	*Radial-Basis Functions*
SBS	*Sequential Backward Search*
SFS	*Sequential Forward Search*
SGBDs	Sistemas de Gerenciamento de Banco de Dados
SRIs	Sistemas de Recuperação de Informação
SCs	Sistemas Classificadores
SiGS	*Silhouette-based Greedy Search*
SiGAS	*Silhouette-based Genetic Algorithm Search*
SiSFS	*Silhouette-based Sequential Forward Search*
StARMiner	*Statistical Association Rule Miner*
SVM	*Support Vector Machine*
trd	taxa de redução de dimensionalidade

Lista de figuras

Lista de Tabelas

Lista de Algoritmos

1 Introdução

Neste capítulo apresenta-se uma visão geral do escopo desta tese, enunciam-se as principais contribuições realizadas por meio dos métodos desenvolvidos e expõem-se os principais resultados alcançados, bem como a organização deste documento.

1.1 Considerações iniciais

As tecnologias de aquisição, comunicação e armazenamento de dados evoluíram além da capacidade humana de assimilação de informação. No domínio médico, por razões de legalidade e valor intrínseco, acervos volumosos de dados digitais têm sido acumulados. Boa parte destes dados são complexos, o que dificulta a aplicação direta de técnicas de Sistemas de Gerenciamento de Banco de Dados (SGBDs), Sistemas de Recuperação de Informação (SRIs) e Sistemas Classificadores (SCs). Assim, o desenvolvimento de técnicas efetivas de consulta, análise e mineração de conhecimento com base nestes dados tem se tornado premente.

A exploração e análise automatizada de dados complexos, tais como imagens médicas, sons e vídeos, são fundamentadas em representações sintáticas que buscam capturar a semântica dos objetos. Representações sintáticas seguem, normalmente, o formato de vetores de características numéricos, gerados por métodos de processamento denominados extratores de características (ou atributos), que estimam valores para propriedades inerentes dos objetos.

No domínio de imagens médicas, em consequência da riqueza semântica e de normal-

mente haver uma grande variação dos aspectos visuais associados a uma mesma patologia, frequentemente torna-se necessária a aplicação de múltiplos extratores [2, 27, 87, 103], o que resulta em representações de alta dimensionalidade, contendo características correlacionadas, redundantes e irrelevantes. Neste cenário, a "maldição da dimensionalidade" (*curse of dimensionality*) [13, 15, 55, 61] – termo utilizado para sintetizar as dificuldades encontradas em espaços de muitas dimensões – degrada o desempenho dos algoritmos de indexação, exploração e análise de dados. Um outro agravante é o problema de descontinuidade semântica (*semantic gap*), que se refere à disparidade existente entre os vetores de características extraídos e a semântica das imagens [5, 24, 26, 35].

Devido aos efeitos colaterais da maldição da dimensionalidade e ao problema de descontinuidade semântica, os sistemas de apoio ao diagnóstico (*Computer-Aided Diagnosis* (CAD)) com base em imagens médicas, têm se mostrado insuficientes em termos de eficácia. Imagens médicas têm um papel fundamental no diagnóstico de pacientes, planejamento cirúrgico, referência médica e treinamento de radiologistas. Atualmente os hospitais contam com o suporte dos Sistemas de Comunicação e Armazenamento de Imagens (*Picture Archiving and Communication System* (PACS)) [111], que têm proporcionado a coleta e a comunicação de dados referentes a exames médicos, formando repositórios ativos de consulta e de apoio à decisão médica. Contudo, ainda são escassas os métodos computacionais efetivos para o aproveitamento da fonte de conhecimento valiosa e facilmente acessível propiciada pelos PACS. Também, um outro fator positivo para fins de pesquisas é o subsídio dado pelo protocolo DICOM (do inglês, *Digital Imaging and Communication in Medicine*) [41] que permite armazenar descrições textuais, conhecidas como metadados, junto com as imagens.

Os métodos computacionais até então utilizados para acesso a exames médicos são as consultas exatas baseadas nos metadados contidos nos cabeçalhos DICOM, tais como: modalidade de exame, patologia e informações pessoais de paciente e de laudos. Apesar deste tipo de consulta proporcionar uma filtragem dos dados do repositório, o número de casos retornados é indeterminado, podendo ser de zero a dezenas. Além disto, eles

não apresentam qualquer ordem de similaridade, o que dificulta a aplicação direta das consultas exatas na prática clínica.

Consultas por similaridade visual ou *Content-Based Image Retrieval* (CBIR), apresentados no Capítulo 4, têm potencialidades para complementar as consultas por metadados implementadas pelos PACS provendo um ferramental efetivo para acesso a dados médicos [77, 78, 81]. Este arranjo entre consultas exatas e por similaridade permite responder consultas tais como: "retorne as 5 imagens mais similares à radiografia de pulmão do João da Silva". Nesta tarefa, pode-se utilizar uma consulta pelo metadado "`radiografia de pulmão`" para obter as imagens desta categoria, dentre as imagens coletadas pelo PACS. Com base neste resultado, uma consulta por similaridade retorna ao médico uma lista ordenada das radiografias de pulmão mais similares à radiografia do João da Silva.

A combinação entre consulta exata e consulta por similaridade é, normalmente, mais eficaz do que o uso de uma modalidade de consulta individualmente. O resultado final, dado pela consulta por similaridade, acrescenta dois elementos importantes para a tarefa de auxílio ao diagnóstico: ordem de similaridade e controle do número de elementos da resposta. Desta forma, o radiologista pode realizar a análise de casos anteriores, conforme a ordem de similaridade retornada pela consulta, e capturar rapidamente informações e evidências que o apoie ou guie na tomada de decisão.

Uma outra ferramenta importante de auxílio ao diagnóstico médico são os Sistemas Classificadores (SCs). Devido à possível falta de concentração, cansaço por longas jornadas de trabalho ou inexperiência frente a casos raros, detalhes patológicos importantes podem passar despercebidos pelos radiologistas, resultando em erros de diagnóstico. Várias pesquisas indicam que o uso de resultados de SCs efetivos pelos radiologista como uma "segunda opinião" eleva significativamente a taxa de acerto de diagnóstico [28, 42, 83, 89].

Contudo, para que os SCs e os sistemas CBIR sejam úteis na tarefa de auxílio ao diagnóstico, é essencial que eles apresentem alta eficácia e eficiência. Em virtude dos desafios proporcionados pela descontinuidade semântica e dos efeitos da maldição da di-

mensionalidade, estes requerimentos não têm sido alcançados. Um modo promissor para amenizar estes problemas é por meio da escolha das características mais significativas das imagens e, consequentemente, a remoção das características desnecessárias, ou seja, a seleção de características. No contexto de aplicações envolvendo imagens, a seleção de características resulta em dois benefícios: redução da descontinuidade semântica, por meio da escolha das características mais relevantes para a aplicação; e amenização dos efeitos da maldição da dimensionalidade, pela remoção das características desnecessárias. No entanto, os métodos de seleção de características existentes, representados principalmente pelas abordagens *wrappers* (que avaliam os subconjuntos de características candidatos com base no desempenho do algoritmo da aplicação meta) e de filtragem (que avaliam os subconjuntos de características candidatos com base em propriedades intrínsecas dos dados), não são efetivamente aplicáveis às tarefas de *Content-Based Retrieval* (CBR) e classificação, principalmente quando se consideram os efeitos agravantes da maldição da dimensionalidade e da descontinuidade semântica.

A tarefa de seleção de características para aplicações CBR, realizada por meio de métodos de filtragem e métodos *wrapper* clássicos (concebidos para minimizar o erro de classificação), tem apresentado resultados insatisfatórios em termos de aumento na precisão das consultas. Os métodos de filtragem normalmente não contam com critérios que permitem a seleção das características mais relevantes para a tarefa CBR. Os métodos *wrapper* anteriores a este trabalho, além de seu alto custo computacional e da sensibilidade ao fenômeno de super-ajustamento (*overfitting*), são inadequados para seleção de características em tarefas CBR, visto que os modelos empregados na avaliação de características são classificadores.

A tarefa de seleção de características para classificação é tradicionalmente realizada por meio de métodos *wrapper*, que avaliam a qualidade de subconjuntos de características com base na acurácia dos resultados produzidos pelo classificador escolhido. Contudo, em situações de alta dimensionalidade, que normalmente é o caso da análise de imagens médicas por conteúdo, além do alto custo computacional, os métodos *wrapper* têm se

mostrado especialmente propensos a *overfitting*, selecionando características que super-ajustam o modelo de classificação empregado.

1.2 Contribuições e resultados principais

Buscando suprir a carência de métodos efetivos de seleção de características para as aplicações CBR (de consultas por similaridade) e de classificação de imagens (cujos resultados são empregados como "segunda opinião") no campo de apoio ao diagnóstico médico, nesta tese foram analisadas as seguintes hipóteses:

1. que funções de avaliação de *ranking* permitem selecionar as características mais adequadas para as aplicações CBR. Com base nesta hipótese foram desenvolvidos os métodos denominados *wrappers* de CBR, onde busca-se pelas características mais significativas para responder as consultas por similaridade;

2. que existe uma simbiose significativa entre o grau de separabilidade entre classes e o desempenho de métodos de classificação. Com base nesta hipótese foram desenvolvidos os métodos denominados filtragem de máxima distinção, onde busca-se pelas características que provêem distinção máxima entre as classes existentes nos dados;

3. que busca GA leva a resultados de seleção de características superiores aos de busca sequencial, pois esta é menos suscetível a soluções mínimas locais, devido a sua propriedade de busca global baseando-se na representação de múltiplas soluções e na aplicações de operadores probabilísticos. Esta propriedade permite lidar, de modo natural, com o aspecto de interação entre características, discutido no Capítulo 2. Além disso, uma busca GA aplicada ao problema de seleção de características tem complexidade de tempo linear enquanto que os métodos de busca sequenciais apresentam complexidade quadrática [62].

A seguir é apresentado uma breve introdução aos métodos *wrappers* de CBR e filtragem de máxima distinção propostos. Ambos se valem de busca de *Genetic Algorithm* (GA) para procurar pelas características mais relevantes nos respectivos domínios de aplicação.

1.2.1 *Wrappers* de CBR

Os *wrappers* de CBR constituem uma nova abordagem de seleção que busca pelas características mais adequadas para responder consultas por similaridade.

Os *wrappers* de CBR constituem uma nova abordagem de seleção de características, dedicada ao aprimoramento de consultas por similaridade por meio da busca pelo subconjunto de características que provêem os *rankings* (respostas das consultas por similaridade) mais adequados (corretos). Seguindo este raciocínio, foi desenvolvida uma família de métodos de seleção de características, que tem como base um conjunto de funções de avaliação dos *rankings* retornados por sistemas CBR. Esta família de funções de avaliação de *ranking* foi denominada *Fitness coach* (Fc), fazendo referência a funções que atuam como técnico (julgador) da corretude de *rankings*. Conforme conhecido, esta é a primeira aplicação de funções de avaliação de *ranking* para seleção de características.

Os resultados obtidos indicam que a seleção de características com base em funções apropriadas de aferimento da qualidade de *rankings*, tais como a família Fc proposta, levam a resultados de CBR (que efetuam consultas por similaridade) significativamente superiores em eficácia aos proporcionados por métodos *wrapper* clássicos (denominados nesta tese de *wrappers* de classificação) e por métodos de filtragem bem estabelecidos na literatura. Quanto à eficiência computacional, os métodos *wrapper* de CBR propostos têm desempenho comparável aos métodos *wrapper* de classificação e são significativamente mais custosos que os métodos de filtragem. Contudo, dado que a seleção de características é normalmente considerada uma etapa de pré-processamento, isto é, realizada uma única vez, e considerando a cardinalidade do conjunto de características na faixa de centenas, o aspecto de custo computacional não é um impedimento para a aplicações dos *wrappers* de CBR na prática.

1.2.2 Filtragem de máxima distinção

Filtragem de máxima distinção é uma nova abordagem de seleção de características para aplicações de classificação, baseada na busca pelas características que levam a maiores

índices de separabilidade entre classes. Foi considerada a hipótese de que existe um nível de simbiose acentuado entre o grau de separabilidade entre classes e o desempenho de métodos de classificação. Com base nesta hipótese desenvolvemos métodos de seleção de características que buscam pelas dimensões dos dados que levam ao maior índice de separabilidade entre classes. O cálculo deste índice é supervisionado, feito por meio da medida de silhueta simplificada [51].

Além dos resultados obtidos confirmarem a hipótese considerada, os métodos desenvolvidos são de baixo custo computacional e superam em eficácia os métodos *wrapper* de classificação tradicionais, devido principalmente à tendência destes a *overfitting*. Quando comparados aos métodos de filtragem da literatura, os métodos desenvolvidos apresentam custo computacional similares, no entanto selecionam características mais adequadas para a tarefa de classificação. Um outro resultado importante obtido é a constatação da supremacia em termos de eficácia da busca GA comparada aos métodos de busca sequenciais. Este resultado é justificado pelas propriedades da busca GA que dificilmente fica presa em soluções mínimas locais (ou máximas locais, dependendo da abordagem dada ao problema) e lidam com o aspecto de interação entre características. Este resultado também indica a existência de interações entre características em representações de imagens dadas pela combinação dos vetores de características gerados por múltiplos extratores.

1.3 Organização do documento

O restante deste documento é organizado do seguinte modo:

Capítulo 2: expõe o problema da maldição da dimensionalidade e aborda métodos para sua mitigação com foco em seleção de características;

Capítulo 3 : apresenta definições, conceitos e fundamentos de algoritmos genéticos;

Capítulo 4: sumariza os conceitos básicos de consulta por similaridade e classificação de imagens, que são as principais ferramentas empregadas na construção de sistemas CAD e discute sobre as barreiras existentes neste campo de pesquisa;

Capítulo 5: apresenta as contribuições desta tese à area de seleção de características e de apoio ao diagnóstico médico, por meio do aprimoramento de sistemas CBR (de consultas a casos similares) e de sistemas classificadores (que fornecem "segunda opinião").

Capítulo 6: sumariza as conclusões principais desta tese, os resultados alcançados, além de apontar questões para investigação futura;

2 Redução de dimensionalidade e seleção de características

Neste capítulo discute-se a importância da redução de dimensionalidade em tarefas de análise e exploração de dados como forma de aliviar os efeitos da maldição da dimensionalidade. Entre as classes de métodos de redução de dimensionalidade, foca-se em seleção de características, após serem discutidas suas vantagens em relação à transformação de características.

2.1 Considerações iniciais

Conforme introduzido no Capítulo 1, a alta dimensionalidade é um aspecto comum em aplicações de apoio à decisão médica. Isto é devido principalmente à natureza semântica diversa das imagens, que torna imprescindível a aplicação de múltiplos extratores de características na busca de uma representatividade adequada. Contudo, o elevado número de características geradas leva ao fenômeno da maldição da dimensionalidade, que aumenta a complexidade de tarefas de manipulação e análise de dados e, consequentemente, degrada o desempenho dos métodos que executam estas tarefas. Os métodos de redução de dimensionalidade constituem os principais antídotos no combate aos males da alta dimensionalidade.

Conforme ilustrado na Figura 2.1, existem duas amplas classes de métodos de redução de dimensionalidade: seleção de características e transformação de características. A transformação de características mapeia as características de seu espaço original para um

novo espaço de menor dimensionalidade. Nenhuma das dimensões originais são mantidas, reduzindo a compreensibilidade dos resultados. Além disso, os processos de transformação, normalmente, não fazem distinção entre características relevantes e irrelevantes conforme o conceito meta (descrição do fenômeno de interesse, i.e., o que deseja-se aprender), fazendo com que as influências negativas das características irrelevantes reflitam no resultado final. Por outro lado, a seleção de características busca encontrar o subconjunto de características mais relevantes do conjunto de dados original de acordo com um critério de avaliação, sendo eliminadas as características desnecessárias.

Os métodos de seleção de características podem ser classificados, de acordo com o critério de avaliação empregado, em filtragem, *wrapper*, embutido e híbrido. Os aspectos de cada uma destas abordagens são discutidos mais adiante neste capítulo.

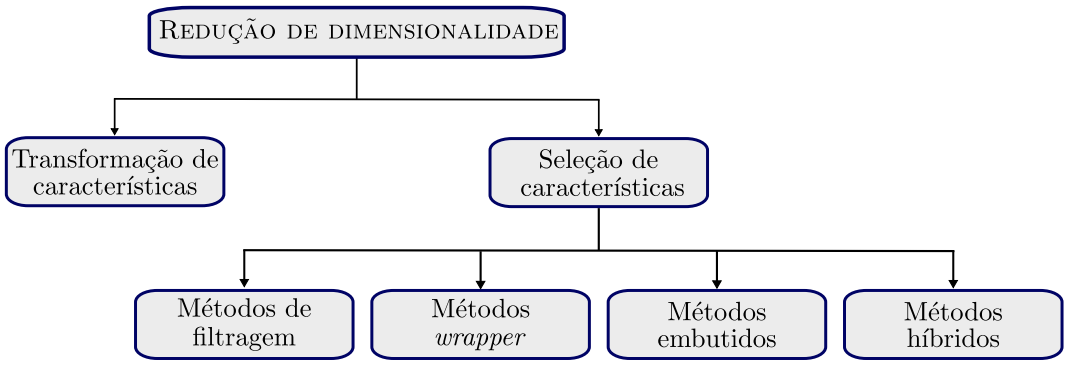


Figura 2.1: Categorias principais de técnicas de redução de dimensionalidade.

No jargão de reconhecimento de padrões, a transformação de características é conhecida pelo nome de extração de características [30, 112]. Nesta tese, o termo extração de características é atribuído somente ao processo de aferimento de aspectos intrínsecos de objetos complexos, tais como, de aspectos visuais de imagens e, portanto, não é considerado sinônimo de transformação de características.

2.2 Maldição da dimensionalidade

É intuitivo pensar que, quanto maior a quantidade de características, mais informações estariam disponíveis para a busca e mineração de dados. No entanto, conforme o número de características cresce, surgem vários fenômenos críticos, tais como: 1) esparsidade

de objetos (ou instâncias) resultando em nivelamento das distâncias entre os mesmos, 2) aumento exponencial do espaço de busca para as tarefas de AM e 3) irrelevâncias, correlações e redundâncias de características. Estes desafios constituem os principais efeitos da maldição da dimensionalidade [13, 15, 55, 61].

O fenômeno dos dados é explicado matematicamente pelo fato da densidade de amostragem de um espaço de m dimensões contendo n objetos ser proporcional a $n^{1/m}$. Assim, mantendo o número de objetos n constante e aumentando a dimensionalidade m, tem-se uma queda exponencial da densidade de amostragem e, consequentemente, o fenômeno de objetos esparsos. Em [15] e [55] é mostrado que o fenômeno de objetos esparsos leva ao nivelamento das distâncias entre os mesmos. Neste caso, é dito que se tem uma indistinguibilidade de vizinhos mais próximos, pois existem muitos objetos com distâncias similares às dos vizinhos mais próximos. Para tornar a situação ainda pior, a busca aos vizinhos mais próximos torna-se mais cara, pois existem muitos objetos fortes candidatos a vizinhos mais próximos, forçando a operação de busca a examinar muitos objetos antes de determinar os verdadeiros vizinhos mais próximos [119]. Além disso, vizinhos mais próximos indistintos são pouco informativos, uma vez que não há diferença significativa entre os vizinhos mais próximos e os outros objetos.

Para ilustrar os fenômenos estudados em [15] e [55], foi realizado um experimento com um dos conjuntos de dados utilizados nesta tese. O experimento teve como base o conjunto *ImageCLEFMed09* (Tabela 5.1), que contém 5000 imagens e uma representação de 1039 dimensões (características). Foram calculadas as distâncias médias entre todos os pares de imagens considerando 2, 4, 8, 16, 32, 64, 128, 256, 512 e 1024 dimensões tomadas aleatoriamente. A Figura 2.2(a) mostra o gráfico obtido. Pode-se observar que a distância mínima e a média se aproximam da distância máxima à medida em que aumenta-se o número de dimensões. Este fenômeno, levado ao extremo, equivale a um nivelamento das distâncias, i.e., um estado no qual as distâncias entre pares de objetos não resultam em diferenças significativas. Em uma consulta por similaridade (Figura 2.2(b)), por exemplo, a distância do elemento de consulta ao vizinho mais próximo seria muito

similar a distância do elemento de consulta ao vizinho mais distante, indicando uma pobre distinção dos objetos (ou instâncias). Neste caso, é dito que a busca aos vizinhos mais próximos é indistinguível, pois existem muitos objetos com distâncias similares [55]. Desta forma, é fundamental a seleção de características no intuito de minimizar os efeitos da maldição da dimensionalidade.

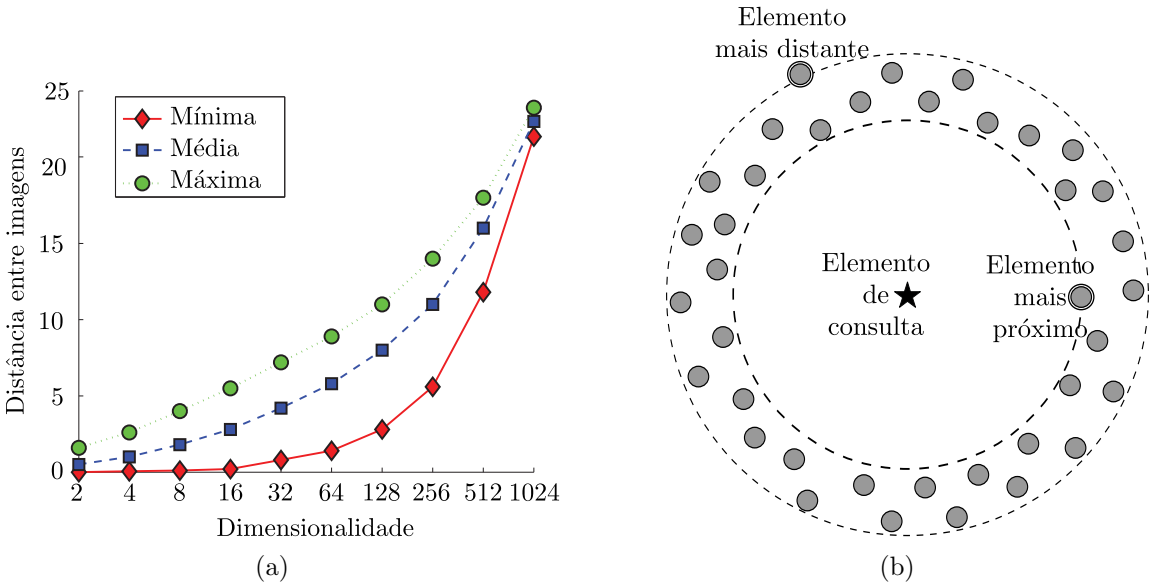


Figura 2.2: Efeitos da maldição da dimensionalidade: (a) Distância mínima, média e máxima entre as imagens do conjunto *ImageCLEFMed09*, considerando variadas dimensionalidades; (b) Efeito ilustrativo da maldição da dimensionalidade em consultas por similaridade.

A maldição da dimensionalidade também degrada o desempenho dos algoritmos de aprendizado de máquina supervisionados do seguinte modo: quanto maior a dimensionalidade dos dados, maior tende a ser a complexidade dos modelos aprendidos com base nos dados de treinamento de modo a minimizar a taxa de erro obtida. Contudo, estes modelos altamente complexos normalmente apresenta o problema de *overfitting*, onde estes super-ajustam aos dados de treinamento e consequentemente apresenta um desempenho insatisfatório sobre os dados de teste. A Figura 2.3 ilustra o aspecto de *overfitting* dos modelos de AM supervisionados em função da alta dimensionalidade dos dados.

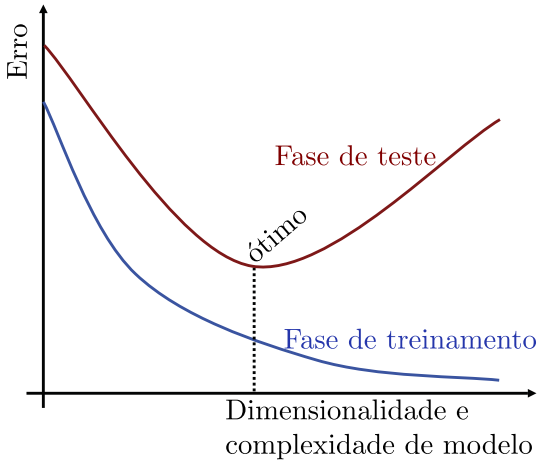


Figura 2.3: *Overfitting* em modelos AM supervisionados função da alta dimensionalidade dos dados.

2.3 Seleção de características

Seleção de características consiste na busca pelas características mais relevantes conforme um dado critério. Seus benefícios palpáveis incluem:

- auxílio na limpeza e compreensibilidade dos dados, possibilitando relacionar as características aos conceitos meta. Por exemplo, associar características de imagens com patologias ou identificar as características mais relevantes para uma determinada tarefa;

- possibilidade de geração de modelos de dados mais simples e mais compreensíveis ao selecionar um subconjunto reduzido das características originais;

- aprimoramento do desempenho dos métodos de mineração, visualização e de consultas aplicados aos dados, em termos de eficiência e eficácia. Os ganhos em eficiência ocorrem em virtude de economia de espaço em memória e de operações computacionais na manipulação dos dados. Os ganhos em eficácia são resultados da remoção de características irrelevantes, ruidosas e correlacionadas, as quais degradam a representatividade das características relevantes;

- redução dos custos, não somente econômicos, associados a cada característica, tais como: sensores físicos, testes médicos, exames e cirurgias invasivas, entre outros.

Isto é, redução dos custos da aplicação alvo;

- redução de tamanho de amostra (número de exemplos de treinamento) necessário em aplicações de aprendizagem de máquina.

Os desafios principais de seleção de características, considerando os efeitos da maldição da dimensionalidade, são:

1. **Espaço de busca:** a cardinalidade do espaço de busca de seleção de características, sem restrições quanto ao número de características desejadas (d), é (2^m - 1) onde m é a dimensionalidade do conjunto de dados considerado. Este fato faz com que uma busca exaustiva seja intratável, mesmo para funções de avaliação de baixo custo computacional e valores moderados de m. A Figura 2.4 ilustra os subconjuntos de características existentes em um espaço de quatro dimensões ($m = 4$). Para $m = 40$, por exemplo, tem-se mais de um trilhão (10^{12}) de subconjuntos de características possíveis, o que levaria mais de 34 anos de execução supondo a avaliação de mil subconjuntos por segundo.

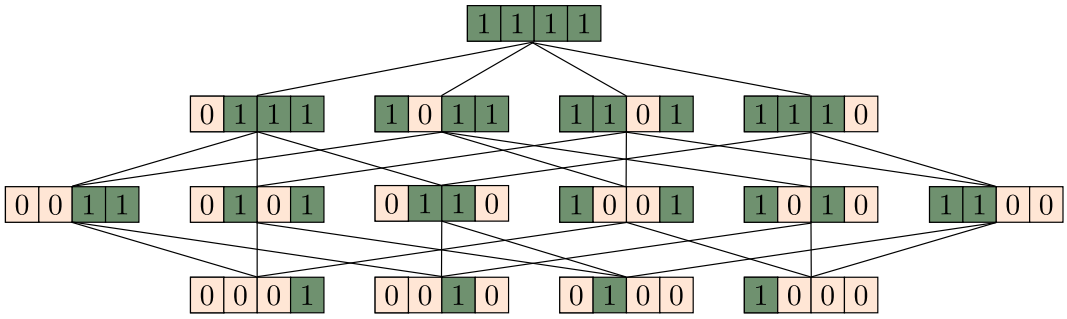


Figura 2.4: Subconjuntos de características possíveis para m=4.

2. **Eliminação de características irrelevantes:** dada uma tarefa de busca ou análise sobre um conjunto de dados de alta dimensionalidade, é provável que muitas características sejam inúteis para o propósito almejado. Características irrelevantes degradam a representatividade daquelas que são relevantes, trazendo sérios impedimentos às aplicações de mineração e de consultas por similaridade.

3. **Eliminação de redundâncias:** em conjuntos de dados de alta dimensionalidade é provável que muitas características contenham a mesma informação, isto é, contêm informações redundantes. Este aspecto é indesejável pelas aplicações de mineração e de consultas por similaridade devido ao aumento de suas complexidades.

4. **Manutenção das características interagentes:** o aspecto de características interagentes, também conhecido como interação de características, consiste de características irrelevantes isoladamente, mas altamente relevantes em conjunto com outras. Deste modo, a remoção de qualquer característica interagente leva a perda de informação das outras características do conjunto de interação. Esta propriedade é denominada irredutibilidade e indica que não se deve avaliar subconjuntos de características interagentes por partes. A Tabela 2.1 mostra um exemplo clássico de interação – a função lógica XOR, que assume valor 1 se A_1 e A_2 assumirem valores diferentes. Observe que, quando A_1 ou A_2 são considerados isoladamente não é possível determinar a valor da função XOR. Várias pesquisas indicam que interação de características é um aspecto comum em aplicações envolvendo dados reais [20, 37, 127].

A_1	A_2	$XOR(A_1, A_2)$
0	0	0
0	1	1
1	0	1
1	1	0

Tabela 2.1: Exemplo de interação de características: função lógicaXOR.

As características interagentes não constituem um problema para as aplicações. Pelo contrário, elas são relevantes na determinação dos conceitos meta e devem ser preservadas pelas operações de seleção de características. Em geral, uma característica é considerada relevante se: 1) ela é fortemente correlacionada ao conceito meta, ou se 2) ela forma com outras características, um subconjunto que é fortemente correlacionado ao conceito meta. Se uma característica é relevante devido à segunda opção, então diz-se que a característica é interagente, i.e., ela interage positivamente com outras características.

A obtenção de um método de seleção de características, que busca pelas características relevantes conforme um conceito meta, pode ser considerada um problema de busca. Para estimar o grau de adequação de características ao conceito meta utiliza-se uma função ou critério de avaliação. Assim, métodos de seleção de características resultam basicamente da combinação de um algoritmo de busca que gera subconjuntos de características candidatos e um procedimento de avaliação destes.

A Figura 2.5 apresenta uma visão geral do ciclo de desenvolvimento de métodos de seleção de características, o qual é composto de duas fases principais: I) a seleção de características em si e II) a avaliação de qualidade do subconjunto de características selecionado – normalmente feita por meio de um algoritmo de mineração ou exploração de dados. Na Fase I, correspondente à seleção de características em si, um algoritmo de busca gera subconjuntos de características candidatos e os envia ao módulo de avaliação (Componente ② da Figura 2.5) que estima a qualidade destes. Caso o critério de parada seja satisfeito, encerra-se o processo de seleção. Caso contrário, os *scores* de avaliação obtidos são passados à estratégia de busca que irá reformular os subconjuntos candidatos e submetê-los novamente ao processo de avaliação. Este ciclo continua até que o critério de parada seja satisfeito. Na Fase II, o subconjunto de características selecionado na Fase I é avaliado com base no resultado produzido pelo algoritmo de aplicação, considerando os dados de um conjunto de teste. Para uma avaliação confiável é essencial que os conjuntos de treinamento e de teste sejam disjuntos. Nas subseções seguintes são apresentadas as principais estratégias de busca e classes de critérios de avaliação empregadas em seleção de características.

2.3.1 Estratégias de busca

Em aplicações reais em que, normalmente, o número de características varia de dezenas a milhares, é necessário o uso de estratégias de busca apropriadas. Na prática, as seguintes estratégias de busca têm sido empregadas:

Ordenação: as características são ordenadas por mérito individual e as primeiras d são

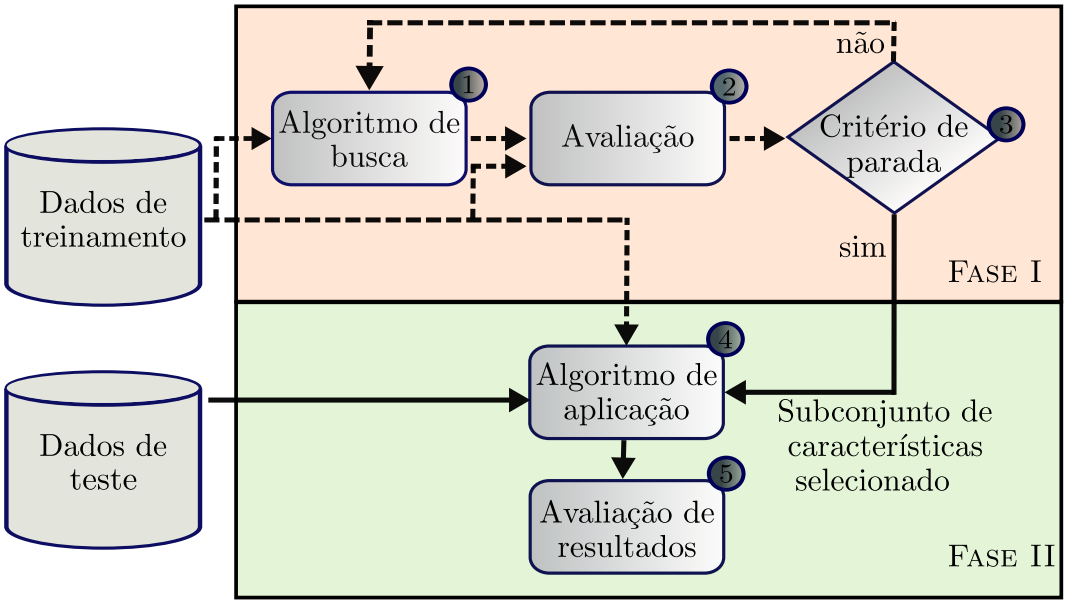


Figura 2.5: Ciclo de desenvolvimento de métodos de seleção de características em duas fases.

selecionadas. Esta abordagem é eficiente computacionalmente pois as avaliações são simples (univariada) e são necessárias somente m avaliações (uma para cada característica). Sua limitação é não considerar a interação entre características, podendo fazer com que o resultado alcançado difira largamente do ótimo em muitas aplicações práticas [20, 37, 127]. Além disso, a escolha de d é um problema por si só.

Sequencial: nesta abordagem as características são adicionadas (ou removidas) ao (do) subconjunto candidato uma a uma. As abordagens mais comumente usadas são: busca sequencial para frente (*Sequential Forward Search* – SFS), onde as características são adicionadas uma a uma ao subconjunto candidato e, busca sequencial para traz (*Sequential Backward Search* – SBS) [112], onde as características são removidas uma a uma do subconjunto candidato. Ambas têm complexidade de pior caso $\mathcal{O}(m^2)$. Contudo, em conjuntos de dados de alta dimensionalidade, SFS é preferida pois permite simplificar o processo de busca, encerrando-o mais cedo, caso nenhuma adição de característica melhore a avaliação.

Metaheurísticas: metaheurísticas locais, tais como: busca tabu (*tabu search*) [108]

e recozimento simulado (*simulated annealing*) [39]; e especialmente globais, tais como: algoritmo genéticos (*genetic algorithms* - GA*s*) [96, 103, 122], colônia de formigas (*ant colony*) [118] e otimização de enxame de partículas (*particle swarm optimization*) [85] têm crescente aplicabilidade em seleção de características. As heurísticas de busca global geralmente provêem resultados superiores aos obtidos pelas estratégias locais por lidarem com o aspecto de interação de características.

Incremental: é um tipo de busca relativamente recente. Embora elas sejam sequenciais no sentido de que é adicionada/removida uma característica por vez, seu comportamento difere amplamente de SFS e SBS [14]. Em cada passo da busca incremental, em vez de avaliar $\mathcal{O}(m)$ candidatos, somente um ou um número constante de candidatos são avaliados. Para obter esse efeito, calcula-se anteriormente, um *ranking* das características usando uma medida de filtragem e, então, um algoritmo de busca sequencial percorre este *ranking* tentando remover/adicionar uma das características em estudo, do/ao subconjunto candidato. A vantagem desta estratégia é que ela possibilita reduzir consideravelmente o número de avaliações *wrapper*, quando comparada aos algoritmos sequenciais.

2.3.2 Classes de métodos de seleção de características

Métodos de seleção de características são usualmente categorizados nos grupos: *wrapper*, filtragem, embutido e híbrido, com base no mecanismo de avaliação de subconjunto empregado (Componente ② da Figura 2.5). Os métodos de cada grupo podem ser supervisionados ou não supervisionados, com exceção dos embutidos que são normalmente supervisionados. Em geral, um subconjunto candidato $\mathbf{A}' \subseteq \mathbf{A} = \{A_1, A_2, \ldots, A_m\}$ é avaliado com base nos dados das colunas da tabela de dados $\mathbf{X}$ indicadas por $\mathbf{A}'$ e na informação de saída desejada, quando esta encontra-se disponível. Estes dados são representados por $\mathbf{X}(\mathbf{A}')$.

Métodos *wrapper*

Os métodos *wrapper* avaliam um subconjunto de características $\mathbf{A}'$ com base no desempenho (normalmente eficácia) de um algoritmo de mineração predeterminado – por exemplo, um classificador, em situações supervisionadas, ou um algoritmo de agrupamento (*clustering*), em situações não supervisionadas – aplicado aos dados de $\mathbf{X}(\mathbf{A}')$. Normalmente, o algoritmo de mineração empregado na avaliação de subconjunto (Componente ② na Figura 2.5) é o mesmo utilizado na aplicação meta (Componente ④ na Figura 2.5). Desta forma, um método *wrapper* busca pelas características de $\mathbf{A}$, mais adequadas para a aplicação meta, pois o algoritmo da aplicação é empregado para avaliar os subconjuntos de características candidatos.

Muitos estudos têm propostos métodos de seleção de características *wrapper* supervisionados [62, 63, 67, 71], sendo alguns deles destinados a maximizar a acurácia do classificador dos k-vizinhos mais próximos (*k-Nearest Neighbor* (kNN)) por meio de busca GA [63, 71, 126]. Os aspectos críticos destes métodos, além do alto custo de execução do algoritmo kNN pela função de avaliação, é que eles são sensíveis ao parâmetro k do classificador kNN [63] e propensos a *overfitting* [65], requerendo a implementação de validação cruzada e realização de testes para encontrar o valor adequado de k, o que aumenta ainda mais o custo computacional. Uma extensão direta dos métodos *wrapper* baseados na minimização do erro do classificador kNN é a sua utilização com base em outros classificadores. Nesta tese, foram empregados os métodos GA-1NN, GA-C4.5, GA-SVM e GA-NB, que utilizam busca GA na tentativa de minimizar a taxa de erro média dos classificadores *1-Nearest Neighbor* (1NN), C4.5, *Support Vector Machine* (SVM) e *Naive Bayes* (NB), respectivamente, como base de comparação aos métodos propostos.

Assim como existem os métodos *wrapper* supervisionados, que têm sido baseados em classificadores, existem os *wrappers* não supervisionados, que são baseados em algoritmos de agrupamento (*clustering*). Um exemplo de um método *wrapper* não supervisionado por ser encontrado em [51]. Este consiste de uma busca sequencial que gera subconjuntos de características candidatos e os passam como parâmetro ao método de agrupamento

k-means. A qualidade dos *clusters* obtidos é então estimada e informada ao procedimento de busca. A meta do processo de seleção é encontrar as características que levam ao maior valor de separabilidade entre *clusters*, de acordo com a medida de silhueta simplificada. A seguir é apresentada a elaboração da medida de silhueta simplificada [51] a partir da definição da medida de silhueta original [56].

Definição 2.1. (Medida de silhueta): [56] Seja $\mathbf{i}$ uma instância pertencente ao *cluster* $\mathbb{A}$ e $a(\mathbf{i})$ a distância média de $\mathbf{i}$ às demais instâncias de $\mathbb{A}$. Seja $\mathbb{C}, \mathbb{C} \neq \mathbb{A}$, um *cluster*. A distância média de $\mathbf{i}$ a todas as instâncias de $\mathbb{C}$ é denotada por $d(\mathbf{i}, \mathbb{C})$. Depois do cálculo de $d(\mathbf{i}, \mathbb{C})$ para todos os *clusters* $\mathbb{C}$, o menor valor é selecionado, i.e., $b(\mathbf{i}) = \min d(\mathbf{i}, \mathbb{C}), \mathbb{C} \neq \mathbb{A}$, que representa a distância da instância $\mathbf{i}$ ao seu *cluster* vizinho mais próximo. A medida de silhueta $\mathfrak{s}(\mathbf{i})$ [56], de uma instância $\mathbf{i}$, é então dada por:

$$\mathfrak{s}(\mathbf{i}) = \frac{b(\mathbf{i}) - a(\mathbf{i})}{\max\{a(\mathbf{i}), b(\mathbf{i})\}} \tag{2.1}$$

É fácil verificar que $-1 \leq \mathfrak{s}(\mathbf{i}) \leq 1$. Quanto maior o valor de $\mathfrak{s}(\mathbf{i})$ mais correto é a atribuição de instância $\mathbf{i}$ para o *cluster* atual, com relação ao princípio da medida de silhueta. A média $\bar{\mathfrak{s}}$ de $\mathfrak{s}(\mathbf{i})$ (Equação 2.2), para toda instância $\mathbf{i}$ pertencente ao conjunto de dados, é usada como critério de avaliação do resultado de agrupamento, onde n é o número de instâncias do conjunto de dados considerado $\mathbf{I}$. Quanto maior o valor de $\bar{\mathfrak{s}}$, mais adequado é o resultado de agrupamento.

$$\bar{\mathfrak{s}} = \frac{\sum_{\forall \mathbf{i} \in \mathbf{I}} \mathfrak{s}(\mathbf{i})}{n} \tag{2.2}$$

A medida de silhueta [56] requer o cálculo de todas as distância entre as n instâncias do conjunto de dados, o que é $\mathcal{O}(n^2)$. Buscando contornar esta limitação, uma versão simplificada desta medida foi proposta em [51]. A medida de silhueta simplificada [51] baseia-se no cálculo das distâncias entre as instâncias e os centróides dos *clusters*. Mais especificamente, a expressão $a(\mathbf{i})$ da Equação 2.1 torna-se a distância da instância $\mathbf{i}$ ao centróide de $\mathbb{A}$. Similarmente, $d(\mathbf{i}, \mathbb{C})$ torna-se a distância de $\mathbf{i}$ ao centróide de $\mathbb{C}$. Estas

simplificações reduzem o custo computacional da medida de silhueta de $\mathcal{O}(n^2)$ para $\mathcal{O}(n)$.

Na figura 2.6 é ilustrado o conceito de silhueta simplificada considerando um espaço bi-dimensional e a função de distância Euclidiana, por simplicidade de visualização. Quanto mais próxima uma instância **i** estiver do centróide de seu *cluster* e quanto mais distante a mesma instância **i** estiver do centróide mais próximo pertencente a um outro *cluster*, maior será o valor da medida de silhueta simplificada de **i**. Isto é, quanto menor o valor de $a(\mathbf{i})$ e quanto maior o valor de $b(\mathbf{i})$, maior é a separabilidade entre os *clusters*, resultando em um valor de silhueta próximo de 1, em condições semi-ótimas.

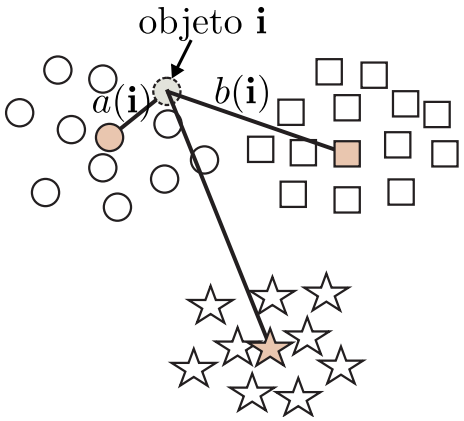


Figura 2.6: Ilustração do conceito de silhueta simplificada.

Nesta tese, a medida de silhueta simplificada foi explorada no desenvolvimento dos métodos de seleção características SiGS e SiGAS.

Métodos de filtragem

Os métodos de filtragem (*filters*) avaliam características individuais ou em subconjuntos baseando-se em propriedades intrínsecas dos dados $\mathbf{X}(\mathbf{A}')$, sem envolver qualquer algoritmo de aplicação. Devido aos cálculos de propriedades intrínsecas, tais como, consistência, medida de informação e correlação serem, normalmente, de baixo custo computacional, os métodos de filtragem são escaláveis mesmo em conjuntos de dados de dimensionalidade elevada.

Dentre os métodos de filtragem mais populares na literatura podem-se citar o *Correlation-based Feature Selection* (CFS) [44], o *Fast Correlation Based-Filter*

(FCBF) [123], o *ReliefF* [92] e o *minimal Relevance Maximal Redundance* (mRMR) [86]. O método *Correlation-based Feature Selection* (CFS) [44] avalia subconjuntos de características usando o cálculo de correlação de Pearson. Quanto menor a correlação entre características e maior suas correlações com a classe, mais adequado é o subconjunto de características. O método *Fast Correlation Based-Filter* (FCBF) [123] emprega medidas de correlação baseadas no cálculo de incerteza simétrica (*symmetrical uncertainty*). Ele seleciona inicialmente todas as características que têm, individualmente, alta correlação com a classe e elimina, uma a uma, as características redundantes, empregando o conceito de *Markov blanket* [123]. O método *ReliefF* [92] estima a qualidade de subconjuntos de características verificando quão bem elas discernem instâncias de classes distintas, próximas umas das outras. O método *minimal Relevance Maximal Redundance* (mRMR) [86] seleciona as características mais correlacionadas com as classes e mais dissimilares das demais, com base nos critérios de máxima dependência, máxima relevância e máxima redundância definidos pelos seus autores.

Implementações dos métodos CFS, FCBF e *ReliefF* encontram-se disponíveis na ferramenta Weka [120]. Uma implementação do método mRMR é disponibilizada por seus autores em `http://www.public.asu.edu/~huanliu/`[1]. Alguns métodos de filtragem não lidam com dados de valores reais ($\mathbb{R}$), requerendo a discretização destes em intervalos. Um método de discretização amplamente conhecido na área de mineração de dados é *Chi2* [66], que consiste em mesclar, iterativamente, intervalos consecutivos de valores de características que levam ao menor decréscimo da estatística χ^2.

Métodos embutidos:

Os métodos embutidos (*embedded*) incorporam a seleção de características ao algoritmo de aplicação, normalmente, um classificador. A relevância de característica é tomada com base na sua utilidade para a otimização da função objetivo do modelo de inferência. Ou seja, a busca pelas características mais relevantes é guiada pelo processo de aprendizagem. Assim, um método embutido provê seleção de características ao mesmo

[1] Acessado pela última vez em 25/03/2011.

tempo em que constrói um modelo de predição. Um exemplo clássico de método embutido é o classificador C4.5 [88], descrito no Capítulo 4, que versa sobre classificação.

Métodos híbridos:

Os métodos híbridos (*hybrid methods*) avaliam as características pelo modo filtragem e pelo modo *wrapper*, explorando a simbiose destes na busca por um melhor desempenho de seleção de características. Normalmente, a avaliação de filtragem é empregada para melhorar a eficiência de um método *wrapper*. Os métodos de seleção de características híbridos são bastante populares atualmente, devido a eles permitirem o aumento da eficiência dos métodos *wrapper*, preservando a sua eficácia. A maior parte dos métodos híbridos de sucesso empregam uma busca global tal como busca GA, refinada por meio de operações de busca local [129, 130].

2.4 Considerações finais

Neste capítulo foram discutidos os efeitos da alta dimensionalidade e a necessidade de técnicas de redução de dimensionalidade para a mitigação de seus males. Foram também discutidos os benefícios e os desafios de seleção de características, seguida pela apresentação dos principais componentes que compõem as técnicas. Na Tabela 2.2 é apresentada uma compilação dos métodos de seleção de características, listando as estratégias de busca comumente empregadas, assim como as principais vantagens e limitações de cada classe de métodos.

No próximo capítulo são apresentados os fundamentos e conceitos básicos de algoritmos genéticos (GAs), que constituem uma técnica de busca empregada com sucesso na seleção de características, devido à sua propriedade de busca por amostragem global de rápida convergência para soluções aproximadamente ótimas. Além disso os GAs raramente ficam presos em soluções mínimas locais e lidam efetivamente com o aspecto de interação entre características.

Modo de avaliação	Estratégia de busca	Vantagens	Desvantagens
Wrapper	– Sequencial – Metaheurísticas	– Seleciona as características mais relevantes para uma dada aplicação	– Alto custo de avaliação de subconjunto – Risco de *overfitting* – Resultado de seleção apresenta viés em favor do algoritmo de aplicação empregado
Filtragem	– *Ranking* – Sequencial – Metaheurísticas	– Baixo custo de avaliação de características – Independe do algoritmo de aplicação	– Características selecionadas podem não ser as mais úteis para uma dada aplicação
Embutido	– *Ranking* – Sequencial	– Simplificação do processo de seleção de características e inferência em um único processo	– Seleção dependente e exclusiva ao método de inferência empregado
Híbrido	– *Ranking* – Sequencial – Metaheurísticas – Incremental	– Mais eficiente que *wrapper* – Herda as vantagens de *wrapper*	– Alta complexidade teórica dos modelos – Herda as desvantagens de *wrapper*

Tabela 2.2: Resumo dos métodos de seleção de características com base no modo de avaliação. Para cada classe de métodos são apresentadas as estratégias de busca possíveis, bem como suas vantagens e limitações.

3 Algoritmos genéticos

Neste capítulo apresentam-se os conceitos principais, os fundamentos básicos e algumas das propriedades dos algoritmos genéticos.

3.1 Considerações iniciais

Algoritmos genéticos, do inglês *Genetic Algorithm* (GA), constituem técnicas de busca/otimização de amplo propósito. A estrutura de soluções potenciais é codificada por uma representação cromossômica e, uma população de cromossomos é evoluída por meio de conceitos básicos de genética (operações de cruzamento e mutação) e seleção natural (operações de seleção), com a finalidade de criar indivíduos mais aptos a cada geração [40, 47].

Usualmente, o processo evolutivo de GAs é encerrado quando as soluções não mais melhoram, ou quando é esgotado o número máximo de gerações preestabelecido. O resultado de saída de um GA é, normalmente, o cromossomo mais apto da população final [40, 47]. Para cada categoria de problema a ser resolvido por GA, deve-se definir uma medida de aptidão (ou função critério). Esta função deve atribuir um *score* para cada cromossomo, diferenciando-os conforme a qualidade (corretude) de solução que cada um representa.

A definição de uma medida de aptidão apropriada ao problema tem um papel essencial na evolução genética [45, 100, 104], pois o *score* calculado é usado no processo de seleção de pares de cromossomos para reprodução e de sobreviventes para gerações consecutivas do ciclo evolutivo. Assim, as maiores probabilidades de reprodução e sobrevivência devem

ser dadas aos cromossomos mais aptos (ou soluções mais adequadas). Devido a sua importância para GAs, as funções de aptidão devem ser feitas "sob medida" para cada categoria problema.

GAs têm a habilidade de lidar eficientemente com grandes espaços de busca e problemas não polinomiais (NPs) [47], além de serem menos propensos a encontrar soluções ótimas locais do que algoritmos de busca determinísticos não exaustivos, tais como as buscas sequencias. Este aspecto de GAs, deriva de seus mecanismos que manipulam múltiplas soluções de modo concorrente, empregando operadores genéticos probabilísticos, promovendo assim, uma eficiente exploração e prospecção do espaço de busca [40, 47].

3.1.1 A inspiração biológica de GAs

Há tempos o homem busca inspiração na natureza para a criação de tecnologias que melhorem sua vivência cotidiana, como por exemplo: aviões inspirados em pássaros, submarinos inspirados em peixes e, sonares inspirados em morcegos. Na comunidade científica há vários estudos sobre métodos e técnicas inspiradas na natureza: redes neuronais inspiradas no funcionamento do cérebro humano [6], sistemas de otimização inspirados no comportamento de colônias de insetos [29], computação evolutiva inspirada na teoria da evolução das espécies [9], dentre outros. Estes campos de pesquisa compõem a área de inteligência artificial (IA), cuja ideia principal é reproduzir artificialmente comportamentos e ações inteligentes observados na natureza, ou realizar tarefas computacionais com base em mecanismos naturais.

GAs constituem umas das técnicas mais difundidas da computação evolutiva. A computação evolutiva estuda os algoritmos evolucionários, que se baseiam na teoria da evolução natural e em interações entre espécies. Os GAs foram criados nos anos 60 pelo pesquisador John Holland que, ao ter acesso aos estudos do biólogo Fisher acerca da evolução natural [36], percebeu um elo nítido entre a biologia e a computação: as máquinas poderiam se adaptar ao meio ambiente, assim como os seres vivos. Conforme sua convicção, a evolução natural era tal como a aprendizagem, i.e., uma forma de adaptação, sendo que a principal diferença entre elas era a duração do processo: várias gerações, ao

invés de uma vida.

O conceito de evolução natural define a natureza como um processo de seleção de seres vivos. Numa determinada população, quando há escassez de recursos, sejam eles comida, espaço, ou outro recurso essencial, os seres mais preparados para a competição se sobressaem e sobrevivem. Isso acontece porque, dentre todas as características imprescindíveis à competição, os seres sobreviventes possuem algumas mais acentuadamente presentes que os outros. Por herança, essas características provavelmente passarão para seus descendentes, e assim, eles também terão grandes chances de saírem vencedores.

Na concepção da genética, um processo evolutivo natural só ocorre se: houver uma população de seres vivos (cromossomos); os cromossomos tiverem a capacidade de reproduzir; houver variedade e a habilidade de sobrevivência estiver associada a essa variedade. Estes fatores tornaram-se essenciais no projeto de algoritmos genéticos. Para Holland, esta semântica da evolução natural e da genética poderia levar as máquinas a evoluírem, assim que fosse desenvolvida uma sintaxe artificial ou um modelo matemático que a suportasse.

Assim, o modo como a evolução foi inicialmente implantada nas máquinas consistiu de partir de um conjunto de possíveis soluções ao acaso e aplicar sobre estas mecanismos inspirados na natureza, desta forma, emergindo um comportamento espontâneo. A evolução de uma população de cromossomos por várias gerações foi transcrita como um processo iterativo de melhoramento das soluções de um problema. As leis da natureza que determinam a sobrevivência dos mais aptos (seleção natural) e promovem a evolução genética foram representadas por operadores artificiais de seleção e de reprodução (cruzamento e mutação), sendo a aptidão de um cromossomo tomada a partir de alguma medida que estima a qualidade da solução que ele representa.

Os GAs foram divulgados à comunidade científica inicialmente em 1975 por meio do livro "*Adaptation in natural and artificial systems*" [49]. Posteriormente, eles tiveram ampla repercussão graças ao livro "*Genetic algorithms in search, optimization and machine learning*" [40]. Atualmente, os GAs dão suporte a várias aplicações computacionais (oti-

mizações em geral, auto-aprendizado, adaptação, previsão, simulação, dentre outras), nas mais variadas áreas do conhecimento, tais como: matemática, biologia, física, química, engenharias, robótica, economia e medicina.

3.1.2 Definições

Conforme em [49], GAs são programas de computador que "evoluem" em um caminho que se assemelha à seleção natural, podendo resolver problemas complexos, até mesmo aqueles que seus criadores não compreendem completamente.

De acordo com [40], os GAs combinam a sobrevivência dos melhores adaptados, com trocas de informações aleatórias e estruturadas, formando um algoritmo computacional com um "faro" inovador de busca. Apesar de aleatórios, os GAs não são uma simples caminhada aleatória. Eles exploram eficientemente informações presentes na população para especular novos pontos no espaço de busca com um aumento esperado de performance.

Conforme em [8], na evolução biológica, a sobrevivência é uma medida de desempenho. Qualquer criatura viva pode ser considerada uma solução estrutural em seu ecossistema. Um GA é um procedimento iterativo que mantém uma população de estruturas candidatas à solução do problema. Durante cada incremento temporal, chamado geração, as estruturas na população corrente são avaliadas por meio de uma medida de desempenho que indica o quão próxima uma estrutura está de ser a solução do problema. Baseada nestas avaliações, uma nova população de soluções candidatas é formada, utilizando três operadores genéticos: seleção, cruzamento e mutação. Cada ponto no espaço de busca do problema é um cromossomo da população, normalmente representado por uma cadeia de símbolos de tamanho fixo.

3.1.3 Características dos GAs

Os GAs diferem da maioria dos procedimentos de busca e otimização em quatro princípios básicos:

1. GAs podem operar tanto em um espaço de soluções codificadas (espaço de genótipos) quanto diretamente no espaço de busca (espaço de fenótipos).

2. GAs operam sobre um ou mais conjuntos de pontos (populações de cromossomos), e não a partir de um ponto isolado, o que os tornam menos propensos a ficarem presos em pontos que são ótimos locais.

3. GAs não necessitam de conhecimentos auxiliares, além da representação das soluções e da estimação da qualidade destas.

4. GAs usam regras de transição probabilísticas e não regras determinísticas.

3.2 Algoritmos genéticos típicos

É denominado de GAs típicos, aqueles que possuem uma única população de cromossomos e otimizam um só objetivo sem empregar busca local. Esta distinção é importante devido a existência de outras classes de GAs, tais como: os meméticos (que empregam busca local junto a busca global), os multiobjetivos (que otimizam simultaneamente um conjunto de objetivos) e os coevolutivos (onde diferentes seres (cromossomos) interagem entre si de vários modos, tais como simbiose, competição, entre outros). Neste texto são apresentados somente os conceitos de GAs típicos, pois eles suportam os desenvolvimentos descritos nesta tese.

O ciclo de execução de um GA típico é mostrado na Figura 3.1. Após eleita uma representação das possíveis soluções de um problema, ou seja, definida a codificação dos cromossomos, gera-se uma população inicial de T_p cromossomos. Os cromossomos desta população são avaliados e, caso estes atinjam o critério de parada, o ciclo é terminado. Caso contrário, um subconjunto destes cromossomos será selecionado e passará por um processo de reprodução (cruzamento e mutação). Os cromossomos descendentes serão avaliados e T_p cromossomos da população "pais + filhos" sobreviverão. Em seguida, é verificado se os cromossomos da população de sobreviventes (população corrente) satisfaz o critério de parada. Caso não satisfaça, os processos de seleção para consequente reprodução, avaliação e seleção de sobreviventes se repetem até que o critério de parada seja atingido. Quando o critério de parada for atingido a(s) melhor(es) solução(ões) será(ão)

apresentada(s). Maiores detalhes das operações deste ciclo são dadas nas subseções a seguir.

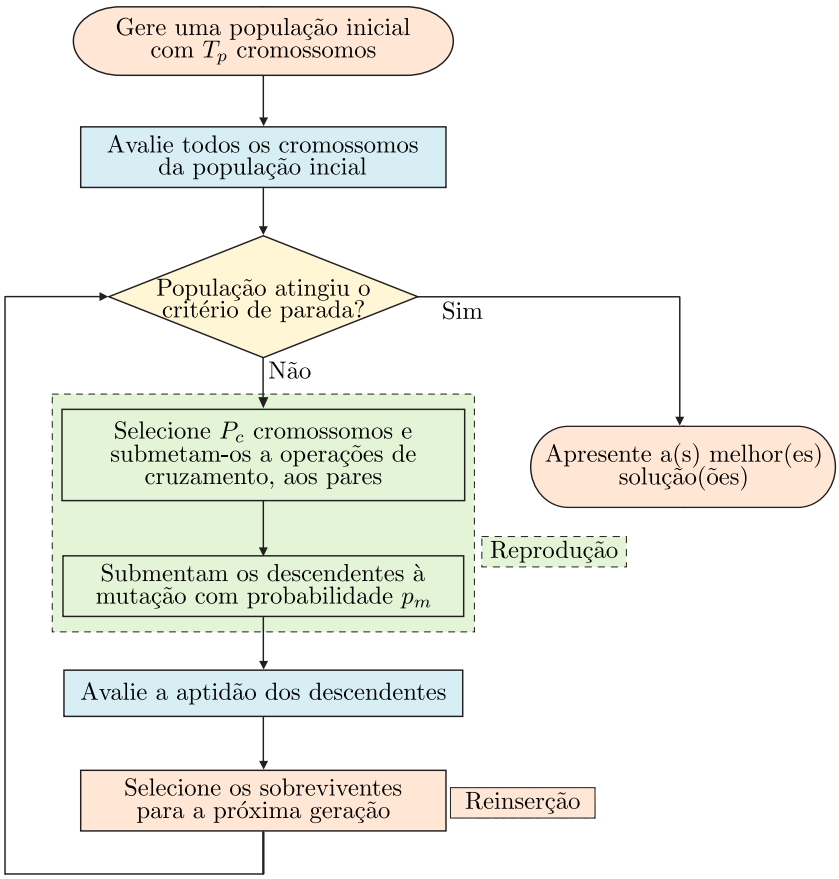


Figura 3.1: Ciclo de execução dos GAs típicos, baseado em [80].

3.2.1 Codificação de cromossomo

Inicialmente é definido o conceito de espaço de busca devido este ser necessário para a apresentação do conceito de codificação dos cromossomos.

Definição 3.1. Espaço de busca: é o conjunto, espaço ou região, que compreende as soluções possíveis de um problema.

O primeiro passo para a aplicação de um GA a um dado problema é eleger um modo de representar cada possível solução (cromossomo) do espaço de busca conforme uma sequência finita de símbolos de um alfabeto. Os primeiros GAs utilizavam exclusivamente representação binária (alfabeto binário). Atualmente representações de cromossomos baseadas em caracteres, números inteiros e reais são bastante utilizadas. A escolha do tipo

de codificação é altamente dependente do problema. Nesta tese empregou-se para seleção de características a codificação binária pois, ela representa as soluções candidatas de modo simples e adequado, facilitando a elaboração das operações genéticas.

3.2.2 População Inicial

Após definida a representação das soluções (cromossomos), uma população inicial de T_p cromossomos é gerada integralmente ou parcialmente de modo aleatório de tal forma que contenha pontos espalhados por todas as regiões do espaço de busca. É importante que a população inicial cubra a maior área possível do espaço de busca, provendo diversidade. Fazendo uma analogia com a natureza, não ocorre evolução sem diversidade, pois é necessário que os cromossomos tenham diferentes características genéticas e, consequentemente, diferentes graus de aptidão, para que possa ocorrer seleção natural.

3.2.3 Medida de Aptidão

A medida de aptidão indica o quão bem adaptado está cada cromossomo da população ao ambiente. Ao longo dos estudos sobre GAs, pesquisas têm mostrado que a especificação de uma medida de aptidão apropriada é crucial para o desempenho das aplicações. É essencial que a medida de aptidão seja bastante representativa, e diferencie na proporção correta, as soluções promissoras das menos promissoras (ou inadequadas) [45, 97, 104]. Se houver pouca precisão na avaliação, soluções promissoras podem ser perdidas durante a execução do GA, que gastará mais tempo explorando soluções pouco promissoras, ou pior, pode ser encontrada uma solução de pouca qualidade. Segundo [73], há vários fatores a serem considerados na elaboração de uma medida de aptidão: característica do problema (maximização *versus* minimização); ambientes determinísticos *versus* indeterminísticos; dinamicidade (o problema se transforma ou evolui no decorrer do tempo); medidas de aptidão alternativas; consideração das restrições do problema e incorporação de múltiplos objetivos.

Normalmente, a medida de aptidão é o componente dos GAs que demanda o maior custo computacional, uma vez que os novos cromossomos, gerados a cada geração ciclo

evolutivo, são avaliados sistematicamente. Pensando em diminuir essa carga computacional, em [47] foram propostos alguns cuidados especiais como: 1) não gerar cromossomos idênticos na população inicial; 2) manter a população com todos os cromossomos distintos entre si, isto é, garantir que a reprodução/evolução não gerará cromossomos idênticos e 3) criar uma memória para os GAs, com o intuito de descartar os cromossomos gerados anteriormente. Na prática, quase sempre, somente o primeiro dos critérios citados é levado em conta na elaboração dos GAs, devido ao custo de manutenção destes cuidados.

Devido aos GAs normalmente partirem de soluções ao acaso, no início da busca os valores de aptidão para os membros da população são bem distribuídos. Quando a busca evolui, valores particulares para cada gene começam a prevalecer. Assim que a variância dos valores de aptidão diminui significativamente, a população converge e, conseqüentemente, não mais evolui, pois já não há o fator imprescindível para a evolução – diversidade. No caso ideal, a população deve convergir para uma solução ótima. Entretanto, em vários problemas reais, não é possível identificar a solução ótima e, conseqüentemente, não se sabe se a população está convergindo para ótimos locais ou para ótimos globais. Análises de convergência e técnicas para a preservação de diversidade são fatores importantíssimos na avaliação e projetos de GAs. Análises de convergência podem ser feitas graficamente por desvio padrão ou por meio de outras técnicas mais sofisticadas de medidas de dispersão.

3.2.4 Seleção

A seleção desempenha o papel da seleção natural na evolução, selecionando preferencialmente, para sobreviver e reproduzir, os cromossomos melhores adaptados ao meio. A seleção é considerada um operador importante na determinação das características de convergência de um GA, sendo vital para estabelecer a pressão seletiva adequada ao ambiente.

Definição 3.2. Pressão seletiva: é o fator que indica o quanto o ambiente é favorável ou desfavorável a um dado cromossomo. Ela modula o grau de privilégio de um cromossomo para sobreviver e reproduzir em detrimento dos demais. A pressão seletiva depende da

medida de aptidão e do operador de seleção adotado. Quanto maior a pressão seletiva, maiores as chances dos cromossomos mais aptos se sobressaírem.

A maneira pela qual os cromossomos são selecionados pode variar, dependendo do operador de seleção utilizado. Os operadores de seleção mais populares são:

- **Seleção estocástica com reposição** - também conhecida como **seleção por roleta**, é o método de seleção padrão dos GAs, proposto originalmente por [49]. A cada cromossomo da população corrente é atribuída uma fatia de uma roleta imaginária, sendo o tamanho desta fatia proporcional à aptidão do cromossomo (Figura 3.2). A cada giro desta roleta é selecionado um cromossomo. Se f_j é a aptidão do cromossomo $\mathcal{C}_j$ na população corrente, a probabilidade P_j do cromossomo $\mathcal{C}_j$ ser selecionado é

$$P_j = \frac{f_j}{\sum_{i=1}^{T_p} f_i}, \tag{3.1}$$

onde T_p é o número de cromossomos na população e f_i é a aptidão do i-ésimo cromossomo.

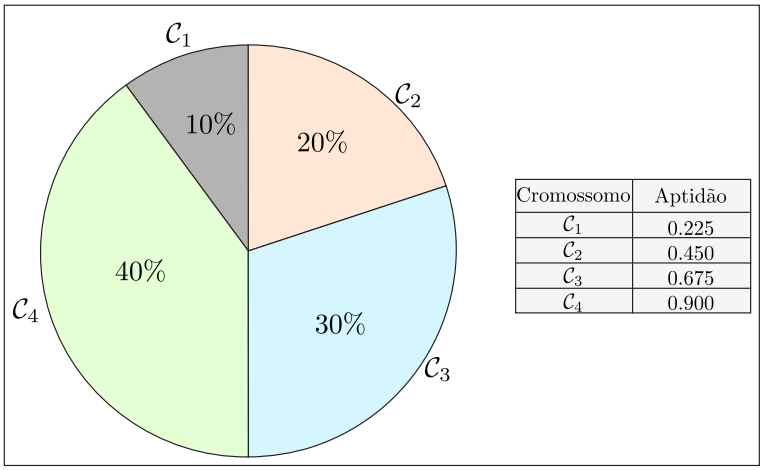


Figura 3.2: Ilustração de uma roleta imaginária utilizada no processo de seleção estocástica com reposição.

- **Seleção por torneio simples:** a ideia é promover um torneio entre um grupo de N ($N \geq 2$) cromossomos aleatoriamente tomados da população. O cromossomo com

o maior valor de aptidão no grupo é selecionado. Neste método, a pressão seletiva pode ser controlada através do tamanho dos grupos.

- **Seleção por torneio estocástico:** análoga ao torneio simples. A única diferença é que os cromossomos dos grupos são selecionados pelo método da roleta, ao invés de serem tomados aleatoriamente.

- **Seleção por truncamento:** um subconjunto dos melhores cromossomos são selecionados, com a mesma probabilidade.

- **Seleção por ordenação:** considerando um problema de maximização, os cromossomos são ordenados pelas suas aptidões, da mais baixa à mais alta. Em seguida, atribui-se a cada cromossomo $\mathcal{C}_j$ uma probabilidade de seleção P_j, tomada de uma distribuição aplicada às posições dos cromossomos no *ranking*. As distribuições mais comuns são, respectivamente, a linear, $P_j = a\ pos(\mathcal{C}_j) + b, a > 0$ e a exponencial, $P_j = a^{b\ pos(\mathcal{C}_j)+c}, a > 0, b > 0$, onde $pos(\mathcal{C}_j)$ é a posição do cromossomo $\mathcal{C}_j$ no *ranking*.

- **Seleção elitista (elitismo):** seleciona diretamente N ($N \geq 1$) cromossomos mais aptos da população corrente. Este operador é normalmente acoplado a outros operadores de seleção, sendo mais empregado para a manutenção dos melhores cromossomos da geração corrente na próxima.

3.2.5 Cruzamento

É um processo inspirado na recombinação biológica, i.e., na troca de material genético entre os pais na geração dos filhos. Pares de cromossomos pais, escolhidos por operadores de seleção, serão submetidos a operações de cruzamento e darão origem a pares de descendentes (filhos). A quantidade de cruzamentos efetuados a cada geração é controlada pelo parâmetro P_c (probabilidade ou taxa de cruzamento). A expectativa é que o cruzamento entre cromossomos bem adaptados gere descendentes cada vez melhores.

O modo como as operações de cruzamento são realizadas depende do domínio e das restrições do problema em questão. As operações de cruzamento mais usuais têm forte inspiração biológica, sendo os filhos formados a partir de trocas diretas de material genético entre os pais. Nesta categoria existem basicamente três tipos de operações de cruzamento: simples, múltiplo e uniforme.

- **Cruzamento simples:** um ponto dos cromossomos, conhecido como ponto de cruzamento, é escolhido aleatoriamente. Ambos os cromossomos pais são cortados neste ponto. A primeira parte do Pai A é concatenada à segunda parte do Pai B, formando um dos filhos (Filho A), e a primeira parte do Pai B é ligada à segunda parte do Pai A, formando o outro filho (Filho B). Um exemplo deste procedimento é ilustrado na Figura 3.3.

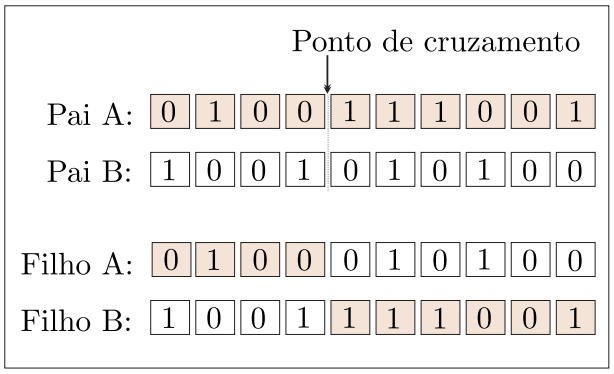


Figura 3.3: Exemplo de cruzamento simples entre o Pai A e o Pai B.

- **Cruzamento múltiplo:** dois ou mais pontos de cruzamento são escolhidos aleatoriamente. As informações genéticas, entre os pontos de corte, são trocadas alternadamente entre os pais. Um exemplo é dado na Figura 3.4.

- **Cruzamento uniforme:** é um tipo de cruzamento múltiplo levado ao extremo, i.e., ao invés de sortear pontos de corte, sorteia-se uma máscara do tamanho do cromossomo, que indica qual cromossomo pai fornecerá cada gene ao primeiro filho. O segundo filho é gerado pelo complemento da máscara. A Figura 3.5 mostra um exemplo de cruzamento uniforme.

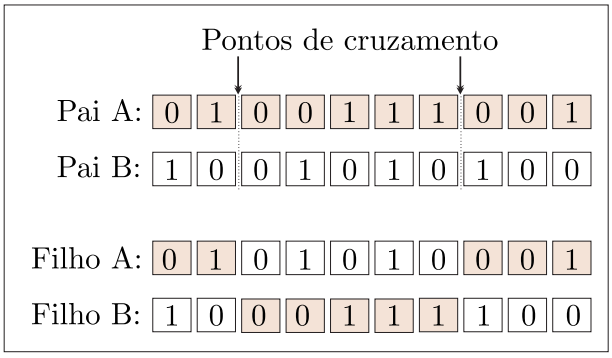


Figura 3.4: Exemplo de cruzamento múltiplo entre o Pai A e o Pai B.

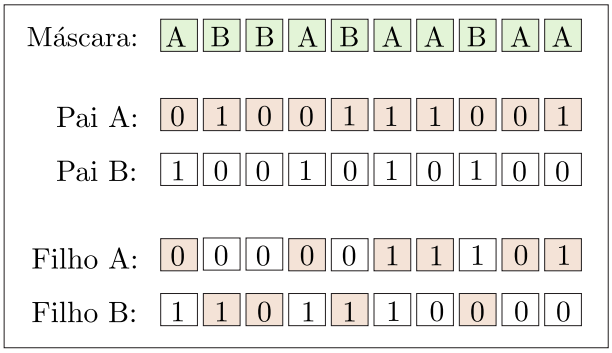


Figura 3.5: Exemplo de cruzamento uniforme entre o Pai A e o Pai B.

Em alguns domínios, a operação de cruzamento não deve gerar genes repetidos. Um exemplo clássico é o problema do Caixeiro Viajante, onde procura-se um trajeto em que o caixeiro passe uma vez em cada cidade, na ordem em que o percurso total seja minimizado. Uma representação de cromossomo natural para este problema é dada por um vetor de números inteiros de m posições, onde cada inteiro corresponde a uma cidade do mapa e a ordem destes indica a ordem de visitação. Dada a restrição do problema, onde cada cidade deve ser visitada uma única vez, o vetor de inteiros não deve ter números repetidos. Consequentemente, a operação de cruzamento não deve gerar cromossomos com genes repetidos. Exemplos de operações de cruzamento que cumprem esta restrição são o *Partially Matched Crossover* (PMX) e o cruzamento cíclico [47].

3.2.6 Mutação

A mutação é uma operação que modifica aleatoriamente alguma(s) característica(s) genética(s) do cromossomo sobre o qual a mesma é aplicada (ver Figura 3.6). Ela é

importante pois permite criar novas características que não existiam na população em análise, introduzindo assim a diversidade genética e assegurando a probabilidade de se chegar a qualquer ponto do espaço de busca [40]. O operador de mutação é aplicado aos cromossomos conforme uma probabilidade de mutação (P_m) geralmente pequena. A probabilidade P_m pode ser aplicada por cromossomo ou por gene.

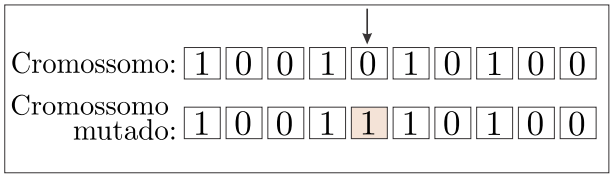


Figura 3.6: Mutação simples.

A operação de mutação, assim como a de cruzamento, deve ser definida de modo a não violar as restrições do problema. Muitos procedimentos de mutação são possíveis, tais como: substituição de um gene por um outro gerado aleatoriamente, perturbação de genes e permutação de genes [40, 47, 75].

3.2.7 Reinserção

Após o processo de reprodução (cruzamento e mutação) e avaliação das aptidões dos filhos, faz-se necessário o uso de um mecanismo de seleção que elegerá os sobreviventes para a próxima geração. Os principais operadores de reinserção são:

- **Reinserção pura:** substitui toda a população pelos filhos. Esta estratégia é normalmente acompanhada de elitismo.

- **Reinserção uniforme:** selecionam-se, a partir de qualquer um dos operadores de seleção tradicionais (Subseção 3.2.4), T_p cromossomos da população total (pais + filhos).

- **Elitismo:** uma parte da população (os melhores pais) é mantida para a próxima geração. Normalmente este procedimento é acompanhado por reinserção pura ou reinserção uniforme.

- **Baseada na aptidão:** também chamada de seleção $(\mu+\lambda)$, a população total (pais e filhos) é ordenada com base nos valores de aptidão e os T_p melhores cromossomos são selecionados.

3.2.8 Condições de Parada

Em problemas de otimização, o ideal é que o GA pare assim que a solução ou o conjunto de soluções ótimas for descoberto [80]. Entretanto, em muitos problemas práticos não se pode afirmar que isto acontece (ou se acontece em tempo viável), até mesmo por não se conhecer as soluções ótimas. Como conseqüência, utilizam-se vários outros critérios de parada como:

- Esgotamento do número máximo de gerações (iterações) pré-estabelecido.
- Esgotamento do tempo máximo de processamento previamente estabelecido.
- Encontro de um cromossomo com aptidão maior ou igual a um limiar pré-definido.
- Estagnação da população ou do(s) melhor(es) cromossomo(s) após um determinado número de gerações.

3.2.9 Parâmetros de Controle

Os GAs típicos têm seu funcionamento baseado em três parâmetros principais: tamanho de população T_p, taxa de cruzamento P_c, probabilidade de mutação P_m. Estes parâmetros têm grande influência no comportamento de um GA, sendo importantes para evitar o problema de convergência prematura

Definição 3.3. Convergência prematura: a população converge prematuramente para um ponto ou um conjunto de pontos que são ótimos locais.

A intuição normalmente seguida na escolha dos parâmetros de controle é a seguinte:

- Uma população muito pequena, implica pouca cobertura do espaço de busca e, conseqüentemente, maiores probabilidades de convergência prematura. Já uma população muito grande possibilita uma ampla cobertura do espaço de busca, prevenindo a convergência prematura. Porém, implica um elevado custo computacional.

- Quanto maior a probabilidade de cruzamento, mais rapidamente novas estruturas serão introduzidas na população. No entanto, se esta for muito alta, estruturas promissoras poderão ser destruídas mais rapidamente que a capacidade da seleção em mantê-las. Assim, normalmente são utilizadas operações de elitismo para garantir que o(s) melhor(es) cromossomo(s) não será(ão) destruído(s) pelas operações de cruzamento.

- Mutações são vitais para a exploração do espaço de busca e evitam a convergência prematura. Entretanto, uma taxa de mutação muito alta torna a busca essencialmente aleatória.

A escolha dos parâmetros de controle dos GAs depende do problema que está sendo tratado, do tamanho e características do espaço de busca, do custo da função de aptidão, entre outros. Muitos autores defendem a hipótese de que estes parâmetros devam ser determinados empiricamente. Outros acreditam que a variação dinâmica destes faz com que os GAs tenham um melhor desempenho, tornando-se menos sujeitos a problemas de convergência.

Na literatura existem vários estudos relativos a especificação de parâmetros de controle [1, 3, 34, 74]. Segundo [74] as técnicas de determinação dos parâmetros de controle podem ser classificadas conforme a Figura 3.7. De acordo com essa classificação, antes da execução o ajuste é feito de modo empírico (experimental). Durante a execução o ajuste pode ser feito dos seguintes modos: determinístico – os valores dos parâmetros são alterados de acordo com alguma regra predeterminada, como por exemplo, em função do número de gerações; adaptativo – de acordo com informações obtidas do processo evolutivo; ou auto-adaptativo – as informações sobre os parâmetros são codificadas dentro dos cromossomos e também reproduzem e evoluem.

3.3 Considerações finais

Neste capítulo foram apresentados e discutidos os principais conceitos relacionados aos GAs típicos, com o objetivo de estabelecer claramente a terminologia e conceituações

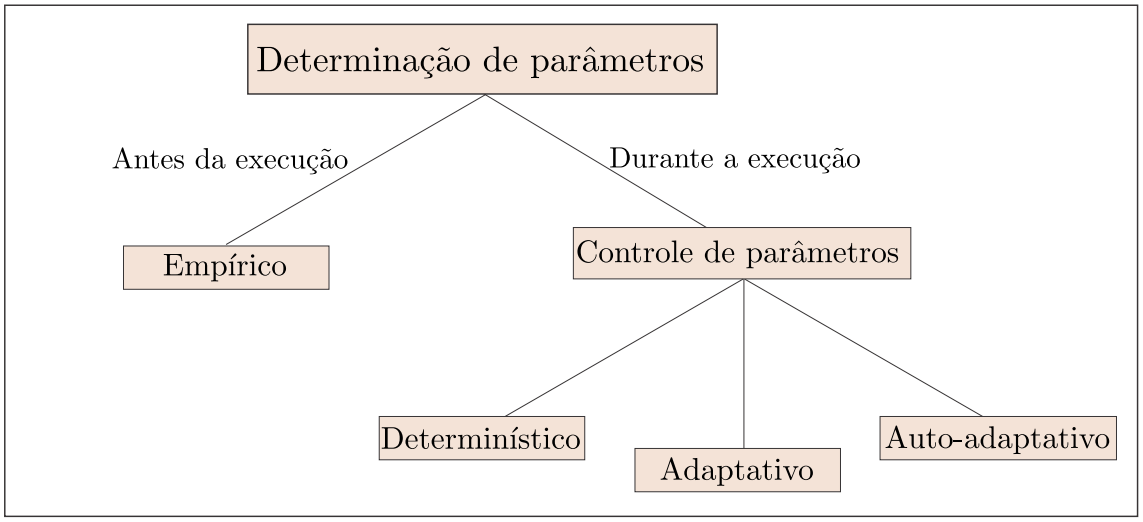


Figura 3.7: Classificação das técnicas de ajuste de parâmetros segundo [74].

utilizadas, bem como evidenciar os aspectos relevantes desta técnica que são destacados na pesquisa descrita nos próximos capítulos.

A otimização é uma ferramenta imprescindível na resolução de problemas complexos. Dentre os métodos de otimização, os GAs se destacam por buscar soluções ótimas sem fazer uso de todo o domínio de soluções candidatas. Isso é possível por causa do uso de técnicas probabilísticas que guiam a população em direção às regiões mais promissoras do espaço de busca.

Os GAs têm sido alvos de vários estudos e têm propiciado aplicações de sucesso em várias áreas do conhecimento, incluindo seleção de características. Neste trabalho, os conceitos de GAs foram explorados no desenvolvimento de métodos de seleção de características eficientes e eficazes na identificação das características de imagens que melhor se aplicam em tarefas de busca por similaridade e classificação no contexto de auxílio ao diagnóstico médico. Buscando cumprir esta meta, foram explorados vários *designs* de funções de aptidão, que mostraram ter um papel fundamental na determinação da qualidade das soluções obtidas.

4 Consultas por similaridade e classificação de imagens

Neste capítulo apresentam-se conceitos e métodos de consulta por similaridade e de classificação de imagens, além de discutir o problema de descontinuidade semântica.

4.1 Considerações iniciais

Os campos de consulta por similaridade e de classificação de imagens podem ser caracterizados como frentes de pesquisa que lidam com várias áreas de conhecimento, tais como: processamento e análise de imagens, reconhecimento de padrões, mineração de dados, recuperação baseada em conteúdo, entre outras. Conforme ilustrado na Figura 4.1, um processo de consulta por similaridade ou de mineração de imagens pode ser dividido em quatro etapas básicas: pré-processamento de imagem, extração de características, integração de dados e mineração ou consultas por similaridade. A etapa de pré-processamento é opcional e tem como objetivo atenuar ruídos e outros aspectos visuais indesejados, ao mesmo tempo em que realça os aspectos importantes para a aplicação. A etapa de extração de características tem a finalidade de gerar representações adequadas das imagens, denominadas de vetores de características, o que fornece a base para a aplicação dos métodos computacionais de apoio à decisão. Na etapa de integração, a representação obtida para cada imagem é associada a dados textuais que descrevem a imagem, o que permitem a realização de consultas mais restritivas e o desenvolvimento de métodos de aprendizado de máquina supervisionados. Em imagens médicas, muitas dessas informações encontram-se

no cabeçalho DICOM (*Digital Imaging and Communication in Medicine*). Por fim, são aplicados os métodos de mineração de dados, que têm como objetivo a extração de conhecimento; ou os métodos de consulta por similaridade, que propiciam a recuperação das imagens do conjunto de dados mais similares a uma dada imagem de consulta. Em geral, a etapa mais desafiadora deste processo é a extração de características que capturem adequadamente a semântica das imagens.

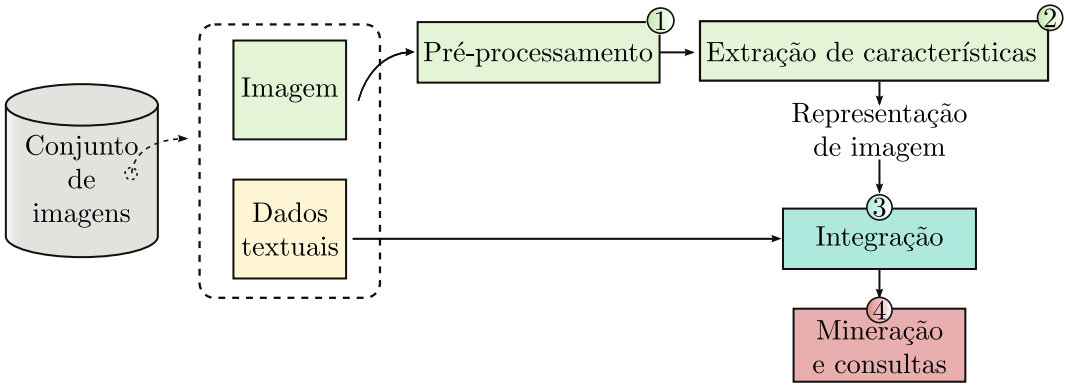


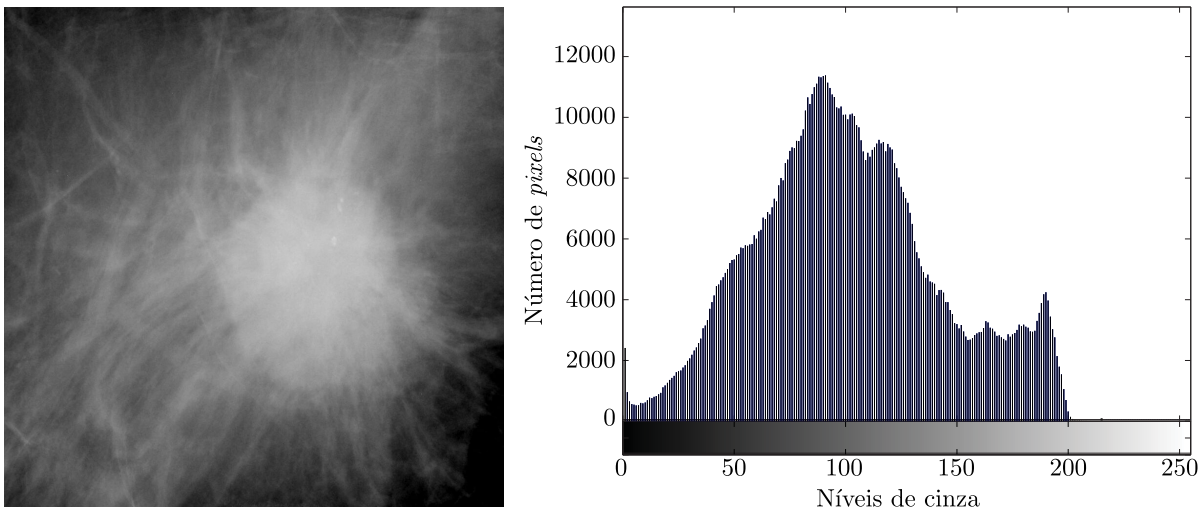
Figura 4.1: Etapas do processo de mineração e consultas por similaridade de imagens. Adaptado de [52].

4.2 Extração de características

As tarefas de consulta por similaridade, análise e mineração de imagens são fundamentadas em representações que sintetizam os conteúdos das imagens. As representações de imagens são denominadas vetores de características (atributos) ou assinaturas. O processo de obtenção de uma representação de imagem é denominado extração de características. Funções de extração de características de imagens normalmente são projetadas para capturar propriedades inerentes das imagens, derivadas principalmente dos aspectos visuais de cor, forma e textura. Um dos principais desafios de consultas por similaridades e análises de imagens por conteúdo em geral é a descontinuidade existente entre as características de baixo nível possíveis de serem extraídas das imagens e os seus conteúdos semânticos associados [5, 24, 26, 35].

4.2.1 Cores

Características baseadas em cores são as mais utilizadas em recuperação por conteúdo, principalmente devido a sua extração ser de baixo custo computacional. Os extratores de características de cor, baseiam-se principalmente em histogramas. O histograma de cores, descrito em [107] é obtido pela quantização do espaço de cores e pela contagem do número de *pixels* que cada cor quantizada possui na imagem. Normalmente, o vetor de características obtido é normalizado pelo número de *pixels* da imagem, de modo a torná-lo invariante às escalas de imagem. A Figura 4.2 apresenta um exemplo de histograma de cores de uma imagem de mamografia quantizada em 256 níveis de cinza. As vantagens de utilizar histogramas normalizados de cores estão na eficiência em termos de sua computação e nas suas propriedades de invariância a transformações de escala, rotação e translação nas imagens.



(a) (b)

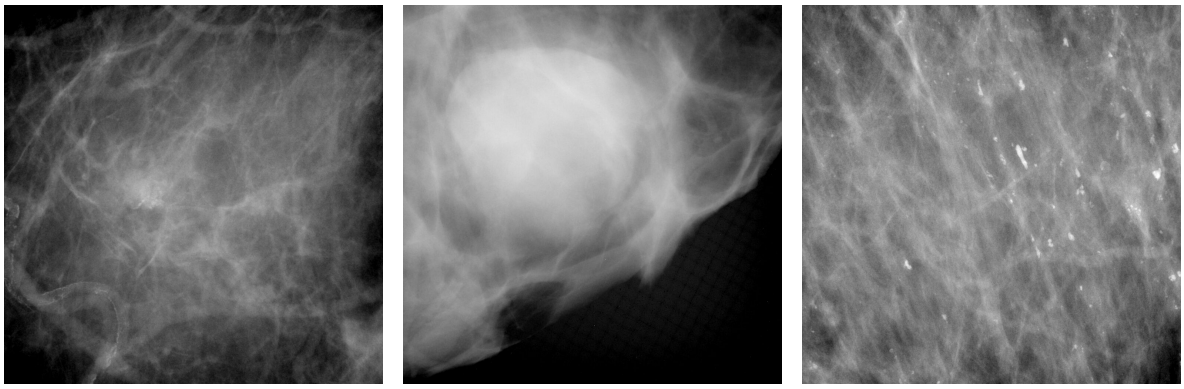
Figura 4.2: Histograma de cores: (a) Imagem de mamografia em 256 níveis de cinza; (b) Histograma de níveis de cinza da imagem (a).

Uma desvantagem do histograma de cores é o fato dele não apresentar informação sobre a distribuição espacial das cores. Diversas técnicas, baseadas no histograma de cores, foram propostas para tratar este problema, entre elas *color coherence vector* [84], *color correlogram* [53] e *color distribution entropy* [106]. Outra desvantagem do histo-

grama de cores é sua alta dimensionalidade. Para reduzir este problema foram propostos os métodos: histograma métrico [116] e *cell histogram* [105]. Em [59] é definido um histograma que explora o conceito de dominância de cores conforme a percepção visual humana. Características globais de cor são combinadas com características espaciais, extraídas considerando uma decomposição *quad-tree* da imagem conforme a distribuição espacial das cores.

4.2.2 Textura

A textura pode ser definida como "o modo como uma pessoa sente uma superfície ao tocá-la, especialmente quanto à maciez ou rugosidade da mesma" [72]. Aplicado a imagens, o termo designa como ocorrem a distribuição de elementos de textura básicos (denominados *textons*) e variações de níveis de cinza. A Figura 4.3 apresenta três diferentes texturas correspondentes a regiões de interesse de imagens de mamografia. É importante destacar que textura é uma das informações mais importantes para classificação de imagens médicas, pois os tecidos normais e os anormais normalmente apresentam propriedades distintas de textura [2, 19]. Entre as técnicas mais importantes para extração de características de textura estão os filtros de Gabor [16], as transformadas de *wavelets* [7, 25], matrizes de co-ocorrência [46], matrizes *run-lengths* [38], as características *Wold* [64] e características Tamura [110].

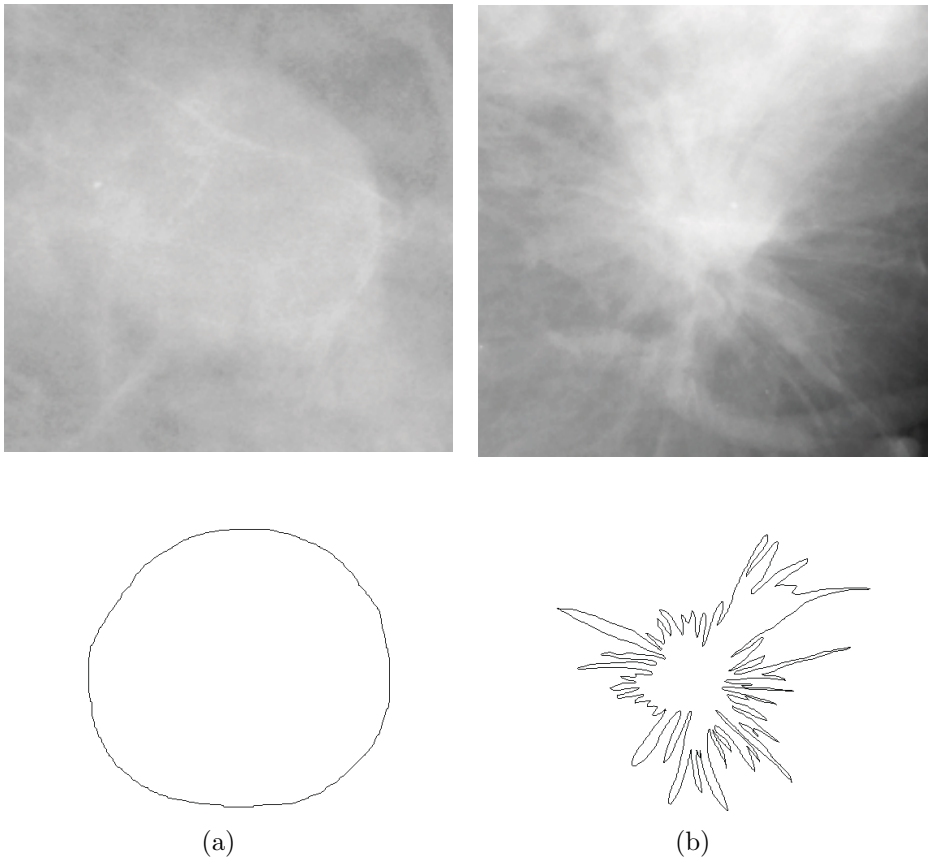


(a) (b) (c)

Figura 4.3: Exemplos de texturas correspondentes a regiões de interesse de mamografia.

4.2.3 Forma

Há várias evidências de que o formato (ou forma) de objetos seja a principal característica explorada pelos humanos no reconhecimento de padrões [79, 82]. Estudos de apoio ao diagnóstico médico relevam que a forma de tumores são de grande importância para classificá-los como benignos ou malignos. Conforme [4], tumores com bordas irregulares têm uma alta probabilidade de serem malignos, enquanto que aqueles com bordas regulares geralmente são benignos. A Figura 4.4 mostra duas imagens de regiões de interesse correspondentes a tumores e seus respectivos contornos.



(a) (b)

Figura 4.4: Massas de tumores e seus respectivos contornos: (a) benigno e (b) maligno. (Fonte [90])

A recuperação de imagens baseada em forma é um dos principais desafios enfrentados pelos sistemas CBR. Isto se deve principalmente à dificuldade de segmentar os objetos de interesse presentes nas imagens. Apesar de inúmeros esforços de pesquisa, a segmentação

automática de imagens ainda é um problema em aberto mesmo em domínios especializados [77]. Depois que as imagens são segmentadas, as características de forma podem ser extraídas com relativa facilidade.

Os métodos de extração de características de forma podem ser classificados em dois grupos [125]: os baseados em contorno e os baseados em região. Os baseados em contorno levam em consideração apenas os contornos dos objetos, partindo da premissa de que os objetos encontram-se segmentados. Já os baseados em região analisam o objeto como um todo.

Extratores de características de forma vão desde simples assinaturas do contorno de objetos contidos nas imagens até descritores mais sofisticados, como os tradicionais descritores de Fourier [121], os momentos de Zernike [57], as saliências de contorno [115] e medidas de dimensão fractal [11]. Com a exceção dos momentos de Zernike que é baseado em região, os demais extratores citados anteriormente são baseados em contorno.

4.3 Consultas por similaridade

Em geral, as consultas tradicionais de SGBDs manipulam dados numéricos, alfanuméricos e textos curtos baseando-se em operadores de igualdade ($=$ e $\neq$) e de ordem total ($<, \leq, >, \geq$). No entanto, para dados multimídia que são de natureza complexa, as consultas clássicas de SGBDs têm pouca utilidade, pois objetos complexos raramente são iguais e não possuem ordem total. Deste modo, as operações de consulta por similaridade são as mais desejadas para estes dados. Após a extração de características de um conjunto de objetos complexos (tais como imagens) e a escolha de uma medida de similaridade (ou função de distância) apropriada, as características extraídas passam a representar cada imagem como um ponto em um espaço m-dimensional, onde m é a quantidade de características. Uma consulta por similaridade pode ser definida por um elemento de consulta e uma restrição baseada na similaridade (distância) ao elemento de consulta. A seguir são apresentados os dois tipos principais de consulta por similaridade.

4.3.1 Consulta por abrangência

Seja $\mathbb{S}$ um domínio de dados. Uma consulta por abrangência (*range query* – *RQ*) recupera todo elemento $\mathbf{e}$ de um conjunto de dados $\mathbf{S} \subseteq \mathbb{S}$ que se encontra a até uma distância (dissimilaridade) máxima r do elemento de consulta $\mathbf{q} \in \mathbb{S}$. Formalmente:

$$RQ(\mathbf{q}, r) = \{\mathbf{e} \in \mathbf{S} | d(\mathbf{e}, \mathbf{q}) \leq r\} \tag{4.1}$$

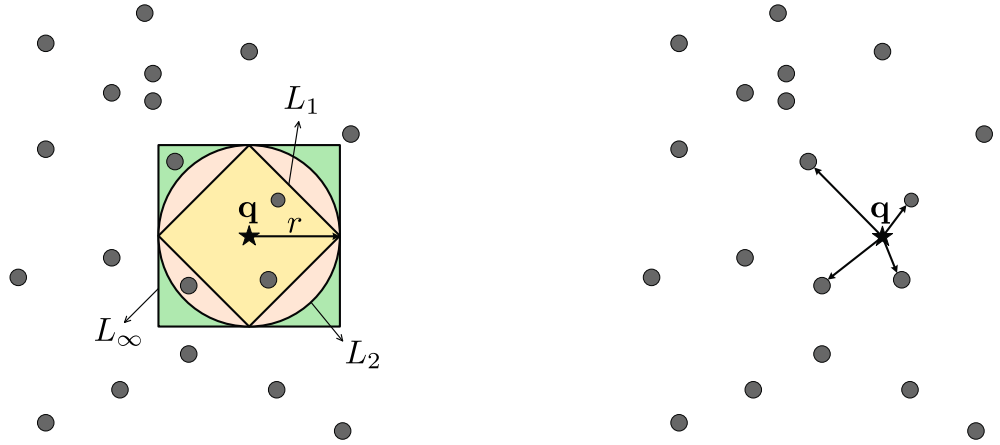
Opcionalmente, os elementos do resultado podem ser retornados ordenados em relação à distância do elemento de consulta $\mathbf{q}$. É importante notar que o elemento $\mathbf{q}$ não precisa fazer parte da coleção de elementos $\mathbf{S}$ que será consultada, porém ele deve pertencer ao espaço m-dimensional. Quando o raio de consulta é nulo ($r = 0$), a consulta por abrangência é chamada consulta pontual (*point query* ou *exact match*). A Figura 4.5(a) ilustra uma consulta por abrangência considerando um espaço de características bidimensional e as funções de distância L_1, L_2 e L_∞ (pertencentes à denominada família L_p), onde as regiões de cobertura para o raio r são:

- L_1: um quadrado de lado $r\sqrt{2}$;
- L_2: um círculo de raio r;
- L_∞: um quadrado de raio $2r$.

4.3.2 Consulta aos *k*-vizinhos mais próximos

Em muitas ocasiões é difícil determinar um raio de busca r sem um prévio conhecimento da distribuição do conjunto de dados e da função de distância. Além disso, uma escolha inadequada pode retornar nenhum, poucos ou uma quantidade demasiada de elementos.

Uma outra opção de restrição de uma consulta por similaridade é informar a quantidade de elementos desejados na resposta. Uma consulta aos k-vizinhos mais próximos (*k-Nearest Neighbor Query* (kNNQ)) recupera os k elementos do conjunto de dados $\mathbf{S} \subseteq \mathbb{S}$ mais similares (próximos) ao elemento de consulta $\mathbf{q} \in \mathbb{S}$. Formalmente:



(a) (b)

Figura 4.5: Tipos de consultas por similaridade: (a) consultas por abrangência, considerando as funções de distância L_1, L_2 e L_∞; (b) consulta kNN para $k = 4$ considerando a distância L_2 (Euclidiana).

$$kNNQ(\mathbf{q}, k) = \{\mathbf{S}' \subseteq \mathbf{S}; |\mathbf{S}'| = k \quad \wedge \quad \forall \mathbf{x} \in \mathbf{S}', \mathbf{y} \in \mathbf{S} - \mathbf{S}' : d(\mathbf{q}, \mathbf{x}) \leq d(\mathbf{q}, \mathbf{y})\} \tag{4.2}$$

A Figura 4.5(b) ilustra uma consulta aos k-vizinhos mais próximos em um espaço euclidiano bidimensional, com $k = 4$. Em [54] é apresentada uma revisão de técnicas de execução de consultas aos k-vizinhos mais próximos.

4.3.3 Estruturas de indexação de consultas por similaridade

Para a realização eficiente de consultas por similaridade é necessário armazenar as características extraídas em métodos de acesso apropriados. O objetivo dos métodos de acesso é realizar a poda de elementos e sub-árvores (conjuntos de elementos) que não fazem parte do conjunto resposta da consulta. Desta maneira, a quantidade de cálculos de distância e, possivelmente, de acesso a disco pode ser reduzida, proporcionando mais eficiência na resposta das consultas.

Entre os principais métodos de acesso para dados multidimensionais destacam-se os métodos baseados na *R-tree* [43] e para dados métricos destacam-se os métodos *M-tree* [21] e *Slim-tree* [117]. Revisões de métodos de acesso multidimensionais e métricos são apre-

sentadas em [93] e [124].

4.3.4 Aprimoramento de consultas por similaridade

O aprimoramento de técnicas de consulta por similaridade em conjuntos de imagens pode ser alcançado por meio de quatro abordagens básicas: 1) composição de descritores [17, 18, 114, 115]; 2) realimentação de relevância [10, 12, 23, 50, 68, 109, 128]; 3) aprendizagem de funções de similaridade entre imagens [113]; e seleção de características [103, 104].

A composição de descritores por meio de testes empíricos, que busca encontrar a função de distância mais adequada a um dado vetor de características, tem se tornado um procedimento padrão na área de CBIR [17, 18, 114, 115]. No entanto, devido à não redução de dimensionalidade, o descritor gerado pode ter desempenho insatisfatório em situações envolvendo vetores de características de alta dimensionalidade, que normalmente apresentam características redundantes e irrelevantes.

Realimentação de relevância é uma das abordagens mais conhecidas e eficazes de refinamento de consultas por similaridade [10, 12, 23, 50, 68, 98, 99, 101, 109, 128]. A ideia principal é usar o *feedback* fornecido pelo usuário acerca da relevância dos documentos previamente recuperados com o intuito de derivar a intenção do usuário de modo que sejam aprimoradas as respostas de consultas futuras. Contudo, a interação de realimentação – onde é necessário opinar explicitamente sobre a relevância os documentos recuperados – não é bem aceita por usuários [60]. Outro fator que deve ser considerado é o tempo de resposta da técnica de realimentação, dado que este é um processo *online*, no qual o usuário fornece *feedback* ao sistema e aguarda por uma resposta, supostamente mais precisa que a anterior.

A aprendizagem de funções de similaridade foi abordada em [113]. O propósito é descobrir uma função que combine os *scores* de similaridade dados por múltiplos descritores de imagem para gerar resultados de similaridade mais adequados a um dado domínio. O método proposto em [113] implementa esta abordagem por meio de programação genética, empregando funções de avaliação de *ranking* como critério.

Embora a seleção de características tenha grande potencialidade na otimização de consultas por similaridade ao amenizar os efeitos da maldição da dimensionalidade e reduzir a descontinuidade semântica, conforme discutido na Seção 1.1 do Capítulo 1, não existem seletores bem estabelecidos na comunidade CBIR. Aplicações de seleção de características em CBIR têm sido realizadas principalmente por meio de métodos de filtragem [31, 32] e métodos *wrapper* projetados com base em classificadores [71], que não selecionam as características mais relevantes para a execução de consultas por similaridade.

4.3.5 Avaliação de desempenho

Sistemas de consulta por similaridade necessitam ser avaliados em termos de eficiência computacional e eficácia. A eficiência computacional é normalmente avaliada em termos de consumo de memória e tempo de processamento necessário para responder consultas por similaridade. Para avaliação de eficácia, tornou-se um padrão o emprego dos gráficos de precisão e revocação (P&R) [10].

Gráficos de precisão e revocação

Seja **I** um conjunto de imagens, também denominado coleção de referência, sobre o qual são executadas consultas por similaridade. Considere que, dada uma consulta **q**, seja conhecido o conjunto de imagens relevantes (Rel). Considere também um sistema de busca que processa a consulta **q** e retorne um conjunto (Rec) contendo as imagens mais similares a **q**. A intersecção dos conjuntos Rel e Rec ($Rel \cap Rec$), compreende os elementos relevantes à consulta **q** que foram recuperados pela operação de consulta. Seja $|Rel|$, $|Rec|$ e $|Rel \cap Rec|$ as cardinalidades dos conjuntos Rel, Rec e $Rel \cap Rec$, respectivamente. Na Figura 4.6 é ilustrada a organização hipotética de tais conjuntos. As medidas precisão e revocação são definidas do seguinte modo:

Precisão: proporção entre o número de imagens relevantes recuperadas ($|Rel \cap Rec|$) e o número de imagens recuperadas ($|Rec|$):

$$\text{Precisão} = \frac{|Rel \cap Rec|}{|Rec|} \tag{4.3}$$

Revocação: proporção entre o número de imagens relevantes recuperadas ($|Rel \cap Rec|$) e o número de imagens relevantes ($|Rel|$):

$$\text{Revocação} = \frac{|Rel \cap Rec|}{|Rel|} \tag{4.4}$$

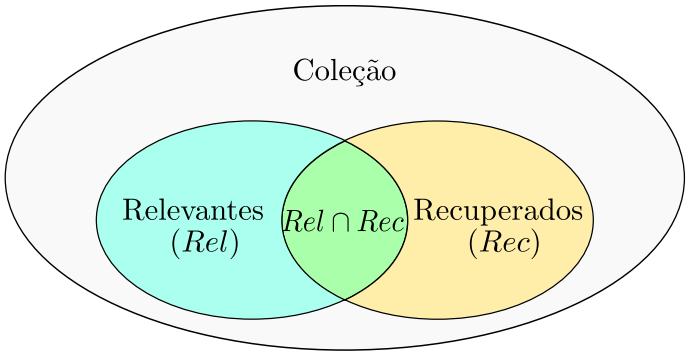


Figura 4.6: Organização em subconjuntos de uma coleção de referência, em termos de documentos recuperados e documentos relevantes para uma dada consulta.

Com base na Equação 4.4 pode-se verificar que a medida de revocação é monotônica crescente em relação a $|Rec|$, pois ela retorna o valor máximo quando todas as imagens da coleção são recuperadas. Já a medida de precisão não tem um comportamento bem definido. Porém, na prática, o valor de precisão tende a diminuir na medida em que a cardinalidade de *Rec* aumenta, pois as imagens são recuperadas em ordem de similaridade e espera-se que, quanto mais próximo ao topo do *ranking*, maior a proporção de imagens relevantes. Devido a estes aspectos das medidas de precisão e revocação, ao invés de se usar valores únicos de precisão e revocação como indicadores de eficácia, utiliza-se um gráfico que ilustra vários valores de precisão e revocação.

Gráficos de precisão e revocação são construídos considerando que a operação de consulta provê um *ranking* (ordenação) das imagens recuperadas *Rec* conforme suas similaridades com relação à imagem de consulta **q**. A Figura 4.7 mostra um *ranking* hipotético em resposta a uma consulta por similaridade, que será empregado para ilustrar a construção

de um gráfico de Precisão e Revocação (P&R). Neste *ranking*, as imagens relevantes retornadas são marcadas com •. Considere também que o conjunto das imagens relevantes para essa consulta, conforme a coleção de referência empregada, seja dada por:

$$Rel = \{\mathbf{i}_5, \mathbf{i}_{13}, \mathbf{i}_{17}, \mathbf{i}_{20}, \mathbf{i}_{31}, \mathbf{i}_{36}, \mathbf{i}_{42}, \mathbf{i}_{47}, \mathbf{i}_{55}, \mathbf{i}_{61}\}, \text{ onde } |Rel| = 10 \tag{4.5}$$

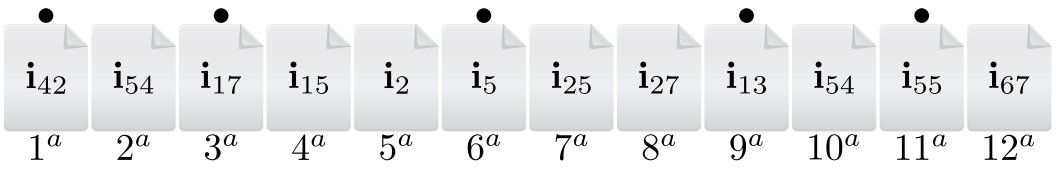


Figura 4.7: *Ranking* de imagens recuperadas. Os marcadores • indicam as imagens relevantes. O número de imagens recuperadas é 12, i.e., $|Rec| = 12$

Examinando o *ranking* das imagens recuperadas (Figura 4.7) verifica-se que o primeiro elemento da lista é relevante. Neste ponto do *ranking*, o valor de revocação é 10%, pois foi recuperado um dos dez elementos relevantes e o valor de precisão é 100%, pois tem-se um elemento analisado e ele é relevante. O próximo elemento relevante é $\mathbf{i}_{17}$, na terceira posição do *ranking*. Neste ponto do *ranking*, o valor de revocação é 20%, pois foram recuperados dois dos dez elementos relevantes e o valor de precisão é aproximadamente 66%, pois há dois elementos relevantes entre os três primeiros retornados. A análise é feita desta maneira para todos os elementos relevantes do *ranking* e, então, é traçado um gráfico com os valores de precisão e revocação obtidos. A Figura 4.8 mostra o gráfico de precisão e revocação para o exemplo recém descrito.

Funções de avaliação de *ranking*

Uma outra medida de avaliação da eficácia de recuperação de imagens é *R-Precision* definida pela Equação 4.6. Ela retorna a porcentagem das *Rel* primeiras imagens recuperadas que são relevantes a uma da consulta $\mathbf{q}$, onde *Rel* corresponde ao número de imagens relevantes no conjunto conjunto de dados de referência. Note-se que *R-Precision*

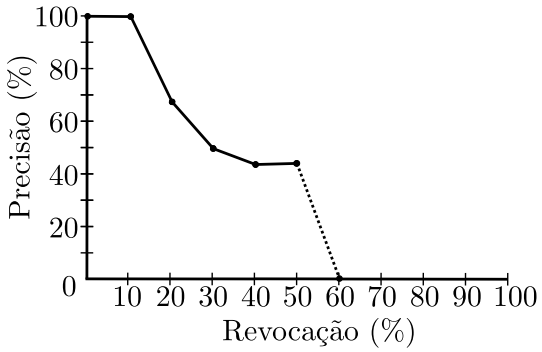


Figura 4.8: Gráfico precisão e revocação para o exemplo da Figura 4.7.

(ou Precisão-R) é um caso especial da medida de precisão (Equação 4.3), considerando o número de imagens recuperadas igual a Rel, que é também igual a medida de revocação da consulta. Esta medida é classificada como uma função de avaliação de *ranking* não baseada em ordem, por não levar em conta o posição de recuperação dos elementos.

$$\text{R-Precision} = \frac{|Rel \cap Rec|}{|Rel|} \tag{4.6}$$

onde Rel é o conjunto das imagens relevantes da coleção, Rec é o conjunto das $|Rel|$ primeiras imagens recuperadas.

Outra classe de medidas, que permite a comparação dos resultados de consultas por similaridade alternativas, são as denominadas funções de avaliação de *ranking* baseadas em ordem, que recebe este nome por considerar a ordem dos elementos retornados em seu cálculo. Uma função de avaliação de *ranking* baseada em ordem que tem apresentado resultados promissores para tarefas de realimentação de relevantes e aprendizagem de funções de similaridade é dada pela Equação 4.7 [69].

$$Fr(\mathcal{L}) = \sum_{\forall \mathbf{i} \in \mathcal{L}} \left(r(\mathbf{i}) \frac{1}{A} \left(\frac{(A-1)}{A} \right)^{(pos(\mathbf{i})-1)} \right) \tag{4.7}$$

onde: $\mathcal{L}$ é o ranking, ou seja, a lista das imagens recuperadas, ordenadas conforme suas similaridades à imagem de consulta; $r(\mathbf{i})$ é uma função que retorna o valor 1 se a imagem $\mathbf{i} \in \mathcal{L}$ sob análise for relevante, caso contrário, ela retorna o valor 0; $pos(\mathbf{i})$ indica a posição da imagem $\mathbf{i}$ no *ranking* $\mathcal{L}$ e $A \geq 2$ é um parâmetro de controle.

4.4 Classificação

A classificação é umas das tarefas mais empregadas em mineração de dados. Um sistema de classificação é utilizado para prever a classe de novos exemplos (objetos) baseando-se em suas características. O objetivo dessa tarefa é criar um modelo computacional com base nas características dos exemplos de treinamento, para prever a classe de novos exemplos. No desenvolvimento de classificadores, os dados disponíveis são divididos em dois conjuntos mutuamente exclusivos: um conjunto de treinamento, usado para a criação do modelo de classificação, e um conjunto de teste, usado para estimar a qualidade do modelo. O conjunto de treinamento fica disponível para o classificador, que analisa as relações entre as características e as classes. Os relacionamentos descobertos a partir desses exemplos (modelo), são então utilizados para prever a classe dos exemplos presentes no conjunto de teste, que fica indisponível ao classificador. Após o classificador prever a classe dos exemplos do conjunto de teste, as classes previstas são então comparadas com as classes reais dos exemplos. Se a classe prevista for igual à real, a previsão foi correta; caso contrário, a previsão foi incorreta. Deste modo, é possível avaliar a acurácia do classificador.

O conhecimento descoberto pelo classificador por meio dos exemplos de treinamento, isto é, o modelo, pode ser representado de várias formas, por exemplo: árvores de decisão [88], redes neurais [48], modelos bayesianos [22] e máquinas de vetores de suporte (SVMs) [94]. Existem também os classificadores que não constroem um modelo para representar o conhecimento descoberto, o que são chamados de classificadores preguiçosos [30]. O exemplo mais conhecido desta categoria é o *k-Nearest Neighbor* (kNN), que será apresentado mais adiante neste capítulo.

O conceito mais difundido para a escolha entre modelos de classificação alternativos é conhecido como Navalha de Occam (Occam' *razor*) [33], que sugere que entre modelos com acurácia similar, o mais simples é preferível. Modelos complexos tendem a possuir um menor poder de generalização, pois estão potencialmente super-ajustados aos dados de treinamento, o que os torna menos eficazes quando utilizados para fazer predições sobre

novos dados. Tal problema é usualmente chamado de *overfitting*.

A seguir são apresentados e discutidos alguns classificadores tradicionais, utilizados nos experimentos desta tese.

4.4.1 Árvores de Decisão

As árvores de decisão classificam padrões com base em uma sequência de testes e decisões. Em geral, uma árvore de decisão representa uma disjunção de conjunções de restrição sobre os valores de característica dos padrões. Cada caminho da raiz da árvore até uma folha corresponde a uma conjunção de testes de característica, e a árvore como um todo corresponde a uma disjunção destas conjunções. Os padrões são classificados seguindo um caminho na árvore da raiz até uma das folhas, a qual provê a classe do padrão. Cada nó interno da árvore corresponde a um teste sobre alguma característica dos dados, e cada ramo descendente a partir de um nó corresponde a uma possibilidade de valor para a característica testada.

Na Figura 4.9 é fornecido um exemplo de árvore de decisão para o problema "jogar tênis", considerando os dados apresentados na Tabela 4.1. A construção de uma árvore de decisão pode ser vista como um particionamento recursivo do conjunto de dados. No nó raiz todas as instância são consideradas e em cada nó filho considera-se somente o conjunto de dados que satisfaz a condição testada. Este processo é repetido recursivamente até que seja satisfeita uma das seguintes condições de parada:

- Todos os dados de um mesmo ramo pertencem a uma mesma classe;
- Não há mais características a serem adicionadas à árvore;
- Não há mais dados de treino.

O aspecto mais importante na construção de árvore de decisão é a escolha da característica corrente de teste, que fará a divisão dos dados. O princípio empregado é o de que árvores simples e compactas são preferíveis às complexas (ideia corroborada pelo conceito de Navalha de Occam [33]). Para este fim, é aplicado um procedimento baseado em um

Tempo	Temperatura	Humidade	Vento	Jogar Tênis
Ensolarado	Alta	Alta	Fraco	Não
Ensolarado	Alta	Alta	Forte	Não
Nublado	Alta	Alta	Fraco	Sim
Chuvoso	Média	Alta	Fraco	Sim
Chuvoso	Baixa	Normal	Fraco	Sim
Chuvoso	Baixa	Normal	Forte	Não
Nublado	Baixa	Normal	Forte	Sim
Ensolarado	Média	Alta	Fraco	Não
Ensolarado	Baixa	Normal	Fraco	Sim
Chuvoso	Média	Normal	Fraco	Sim
Ensolarado	Média	Normal	Forte	Sim
Nublado	Média	Alta	Forte	Sim
Nublado	Alta	Normal	Fraco	Sim
Chuvoso	Média	Alta	Forte	Não

Tabela 4.1: Exemplos de treinamento para o problema "jogar tênis".

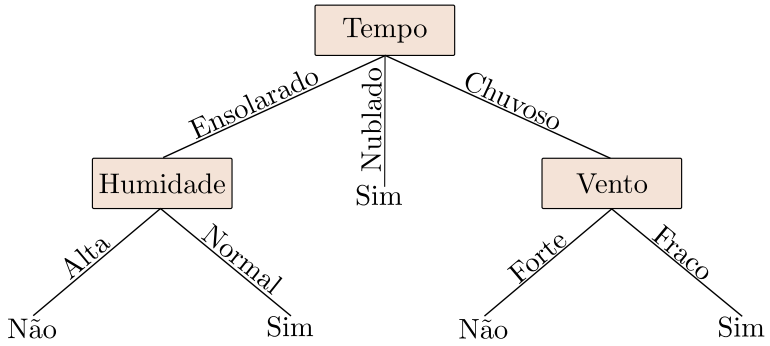


Figura 4.9: Árvore de decisão para o exemplo jogar tênis (Tabela 4.1).

critério de impureza, tal como entropia, que efetue partições resultando em subconjuntos de amostras o mais homogêneas possíveis, em cada ramo da árvore. No decorrer da construção da árvore, uma folha com amostras heterogêneas é substituída por um nó teste que divide o conjunto heterogêneo em subgrupos minimamente heterogêneos, de acordo com o critério de impureza. Em outras palavras, a característica mais informativa em um estágio particular é usada para dividir os dados, pois é a que reduz mais a incerteza.

Como consequência, a operação fundamental de um algoritmo de indução de árvore de decisão é o cálculo de impureza, que determina a divisão a ser realizada em um determinado nó. Existem várias medidas de impureza, todavia, as mais utilizadas são o ganho de informação e a taxa de ganho. Ambas utilizam o conceito de entropia no sentido de

teoria da informação (Entropia de Shannon [95]). Um dado conjunto de padrões $\mathbf{S}$ pode ser descrito em termos de sua distribuição de rótulos de classe, e sua entropia pode ser calculada como:

$$H(\mathbf{S}) = -\sum_{i=1}^{l} P(c_i) \log_2 P(c_i) \tag{4.8}$$

onde $P(c_i)$ corresponde à proporção de padrões em $\mathbf{S}$ pertencente à classe c_i, e l é o número de classes em $\mathbf{S}$.

O ganho de informação $IG(\mathbf{S}, D)$ representa a redução de entropia (incerteza) esperada quando o conjunto $\mathbf{S}$ é dividido com base na característica D, e que pode ser calculado como:

$$IG(\mathbf{S}, D) = H(\mathbf{S}) - H(\mathbf{S}|D) = H(\mathbf{S}) - \sum_{j \in V(D)} \frac{\mathbf{S}_j}{\mathbf{S}} H(\mathbf{S}_j) \tag{4.9}$$

onde $V(D)$ denota os valores possíveis para a característica D e $\mathbf{S}_j$ é o subconjunto de $\mathbf{S}$ para o qual a característica D tem valor j. A característica mais adequada a ser usada como critério de decisão é aquela que resulta no valor máximo de $IG(\mathbf{S}, D)$, pois, maximizando o ganho de informação, minimiza-se o grau de impureza. Contudo, o uso do ganho de informação como critério tem uma desvantagem inerente da entropia, favorecendo características com um alto número de valores possíveis. Para evitar este inconveniente, o ganho de informação deve ser normalizado pela entropia de $\mathbf{S}$ em relação aos valores da característica D, resultando em um outro critério denominado taxa de ganho (*gain ratio*):

$$GainRation(\mathbf{S}, D) = \frac{IG(\mathbf{S}, D)}{-\sum_{j \in V(D)} \frac{|\mathbf{S}_j|}{|\mathbf{S}|} \log_2 \frac{|\mathbf{S}_j|}{|\mathbf{S}|}} \tag{4.10}$$

Um dos classificadores mais conhecidos baseado em árvores de decisão é o C4.5 [88]. O classificador C4.5 pode manipular valores de características contínuos utilizando pontos de corte e introduz medidas para evitar *overfitting* tais como parada da divisão dos nós e poda da árvore. Além disso, ele pode manipular padrões com características ausentes.

4.4.2 Classificadores Bayesianos: *Naive Bayes*

Um classificador bayesiano é um classificador estatístico baseado no teorema de Bayes [112]. O teorema de Bayes é definido do seguinte modo: seja $C = \{c_1, c_2, \ldots, c_l\}$ o conjunto de classes dos dados e $\mathbf{x}$ uma instância de classe desconhecida. Considerando-se que $\mathbf{x}$ pertence a uma das classes do conjunto C, deseja-se determinar $P(c_i|\mathbf{x})$, $1 \leq i \leq l$, ou seja, a probabilidade da classe c_i dada a instância $\mathbf{x}$. O cálculo da probabilidade *a posteriori* da classe c_i condicionada a $\mathbf{x}$, $P(c_i|\mathbf{x})$ é dado pela regra de Bayes:

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})}, \tag{4.11}$$

onde $P(c_i)$ é a probabilidade *a priori* da classe c_i, $P(\mathbf{x})$ é a probabilidade *a priori* de $\mathbf{x}$ e $P(\mathbf{x}|c_i)$ é a probabilidade *a posteriori* de $\mathbf{x}$ condicionada a classe c_i. As probabilidades $P(c_i)$, $P(\mathbf{x})$ e $P(\mathbf{x}|c_i)$ são estimadas a partir das instâncias de treinamento.

Dado um exemplo $\mathbf{x}$ de classe desconhecida, um classificador bayesiano prediz que $\mathbf{x}$ pertence a classe que tem a maior probabilidade *a posteriori* $P(c_i|\mathbf{x})$, i.e., $\arg_{c_i} \max P(c_i|\mathbf{x})$. Considerando $P(\mathbf{x})$ constante para todas as classes tem-se que:

$$P(c_i|\mathbf{x}) = P(\mathbf{x}|c_i)P(c_i) \tag{4.12}$$

O classificador bayesiano mais simples é *Naive Bayes*. Este classificador é denominado ingénuo (*naive*) por assumir que as características são condicionalmente independentes, ou seja, que a informação de um evento não é informativa sobre nenhum outro. Assumindo que as características são condicionalmente independentes dada a classe tem-se que:

$$P(\mathbf{x}|c_i) = \prod_{k=1}^{m} P(x_k|c_i), \tag{4.13}$$

sendo m o número de características dos exemplos e $P(x_k|c_i)$ é estimada dos exemplos de treinamento do seguinte modo:

- Se x_k for categórico, $P(x_k|c_i) = s_{ik}/s_i$, onde s_{ik} é o número de exemplos de treino

da classe c_i que têm o valor x_k para a característica A_k e s_i é o número de exemplos de treino da classe c_i.

- Se a característica A_k for contínua, é assumido que ela possui uma distribuição gaussiana e é calculada a probabilidade como:

$$P(x_k|c_i) = \frac{1}{\sigma_{c_i}\sqrt{2\pi}} e^{-\frac{(x_k-\mu_{c_i})^2}{2\sigma_{c_i}^2}}, \tag{4.14}$$

onde μ_{c_i} e σ_{c_i} são, respectivamente, a média e o desvio padrão dos valores da característica de índice k para os exemplos da classe c_i.

O classificador *Naive Bayes* é simples e, geralmente, apresenta alta precisão preditiva e escalabilidade em grandes bases de dados de alta dimensionalidade [76].

4.4.3 *Support Vector Machines*

As *Support Vector Machine*s (SVMs) foram originalmente formuladas para lidar com problemas de classificação binários (duas classes). Atualmente, existe uma série de técnicas que podem ser empregadas na generalização das SVMs para a resolução de problemas multiclasses. Assim, é apresentado a seguir uma breve introdução às SVMs em duas partes.

Classificação binária

Dado um conjunto de treinamento composto por n amostras, denominadas vetores no contexto das SVMs, pertencentes a duas classes linearmente separáveis, o objetivo é definir um hiperplano que separe os vetores. Entre os muitos hiperplanos possíveis, o hiperplano separador ótimo é o plano que maximiza a margem, ou seja, a distância entre o hiperplano e o vetor mais próximo de cada classe. A Figura 4.10 ilustra este procedimento.

As SVMs lidam com problemas não lineares realizando um mapeamento da forma $\Phi : \mathbf{A} \rightarrow \mathbf{B}$ no qual $\mathbf{A}$ é o espaço de características original do problema e $\mathbf{B}$ o espaço de destino do mapeamento, que deve ter maior dimensionalidade do que $\mathbf{A}$ (veja Figura 4.11).

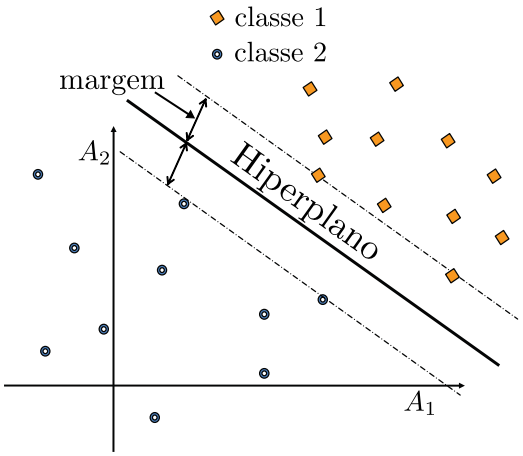


Figura 4.10: Hiperplano de separação SVM de maior margem.

As funções que realizam este tipo de mapeamento são denominadas funções Kernel. Uma escolha apropriada de função Kernel Φ faz com que o conjunto de treinamento **Q** mapeado do espaço de características **A** para **B** seja separável por uma SVM linear (teorema de Cover [48]). Os tipos de funções Kernel mais utilizadas na prática são as polinomiais, gaussianas (*Radial-Basis Functions* (RBFs)) e as sigmoidais.

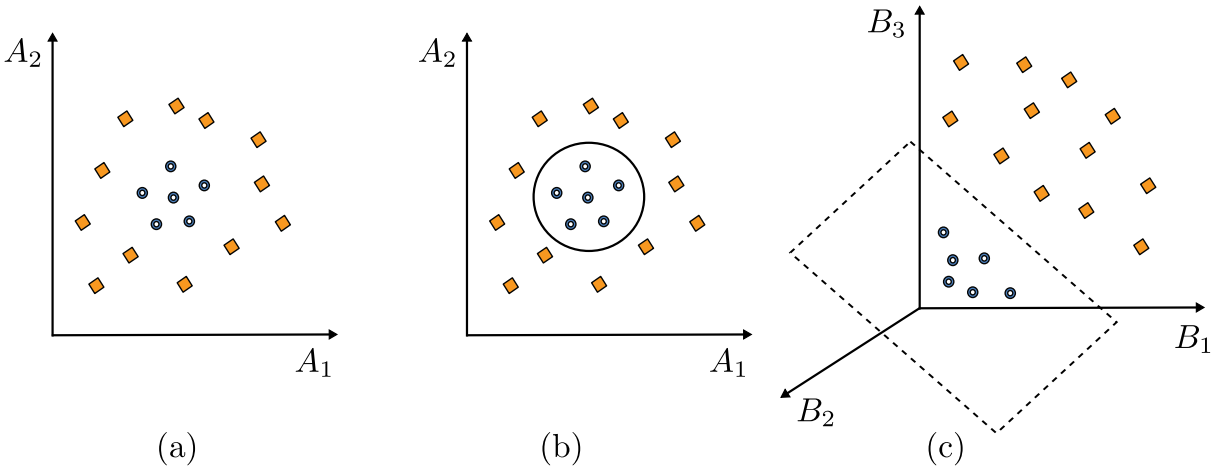


Figura 4.11: Mapeamento de um conjunto de dados não linearmente em um linearmente separável: (a) Conjunto de dados não linear; (b) Fronteira não linear no espaço original; (c) Fronteira linear no espaço transformado.

Classificação multiclasse

Existem basicamente duas abordagens de SVMs multiclasse: a de decomposição do problema multiclasse em vários subproblemas binários e a de reformulação do algoritmo

de treinamento das SVMs em versões multiclasse. Em geral, esse último procedimento leva a algoritmos computacionalmente custosos [52]. Por esse motivo, a estratégia decomposicional é empregada mais frequentemente. Uma revisão a respeito da obtenção de previsões multiclasse com SVMs pode ser encontrada em [70].

4.4.4 Classificadores Preguiçosos: *k-Nearest Neighbor*

Os classificadores apresentados até o momento são caracterizados pelo fato de construírem um modelo de classificação utilizando os dados de treinamento. Normalmente, a construção de modelo demanda um custo computacional considerável, enquanto que a classificação de novos objetos é feita de forma rápida. Tais classificadores são chamados de classificadores apressados (*eager classifiers*). Ao contrário dos classificadores até então apresentados, os classificadores preguiçosos não constroem modelos de classificação na fase de treinamento. Os objetos não rotulados são classificados com base na classe majoritária dos padrões de treinamento que mais se assemelham a eles. Como não é construído um modelo, para cada objeto a ser classificado é analisado todo conjunto de treinamento. Obviamente, este processo é computacionalmente dispendioso, especialmente para conjuntos de treinamento com um elevado número de instâncias. O exemplo mais popular de classificador preguiçoso é o *k-Nearest Neighbor* (kNN) [30].

O classificador kNN é descrito da seguinte forma. Suponha um conjunto $\mathbf{Q}$ de amostras de treinamento. Cada elemento de $\mathbf{Q}$ é uma tupla $(\mathbf{x}, c)$, onde $\mathbf{x}$ é um objeto m dimensional e c é o seu rótulo. Seja $\mathbf{y}$ um novo objeto não rotulado. Com o objetivo de classificar $\mathbf{y}$ calcula-se a distância de $\mathbf{y}$ a todos os objetos de treinamento $\mathbf{Q}$. O rótulo de $\mathbf{y}$ é dado pela classe que ocorre com maior frequência dentre os k objetos mais próximos de $\mathbf{y}$.

Antes do processo de classificação, os valores de características são, normalmente, normalizados para que valores em diferentes escalas não produza *bias* no cálculo de distância [30]. As métricas de normalização mais utilizadas são *standardization* (também conhecida como *z-score*), dada por $z_i = \frac{v_i - \mu(\mathbf{v})}{\sigma(\mathbf{v})}$ e *normalization* dada por $n_i = \frac{v_i - \min(\mathbf{v})}{\max(\mathbf{v}) - \min(\mathbf{v})}$, onde onde v_i é o valor a ser normalizado, $\mu(\mathbf{v})$ e $\sigma(\mathbf{v})$ correspondem à média e ao desvio padrão dos valores em $\mathbf{v}$, onde $\mathbf{v}$ é um vetor de valores das instâncias

do conjunto de dados para uma determinada característica.

4.4.5 Técnicas de amostragem de dados

Buscando melhorar as estimativas de acurácia e diminuir o *bias* em relação aos dados de treinamento podem-se utilizar técnicas de amostragem na construção do modelo, tais como, a amostragem aleatória (*random sampling*) e os métodos de validação cruzada: *k-fold cross-validation* e *leave-one-out*.

***Random sampling*:** consiste em dividir aleatoriamente o conjunto de dados em subconjuntos disjuntos. Por exemplo: 70% das amostras para treinamento e 30% para teste. Este processo pode ser repetido várias vezes buscando uma melhor estimativa média de desempenho de um modelo.

***k-fold cross-validation*:** consiste em dividir o conjunto de dados em k partições mutuamente exclusivas e experimentar o modelo k vezes, utilizando $k-1$ partições para treinamento e uma partição para teste. A taxa de erro é dada pela média dos erros de teste obtidos nas k repetições. Quando a proporção de objetos por classe do conjunto completo é mantida nas partições, este procedimento recebe o nome de *stratified k-fold cross validation*.

***Leave-one-out*:** o modelo é executado N vezes, considerando um conjunto de N amostras. Em cada iteração, $N-1$ amostras são utilizadas para treinamento do modelo e uma amostra é utilizada para teste. A taxa de erro é obtida dividindo o número de erros obtidos nos N testes por N.

4.5 Considerações finais

Neste capítulo foram apresentados fundamentos e métodos de consulta por similaridade e de classificação de imagens. Estes métodos são essenciais no desenvolvimento de sistemas de apoio ao diagnóstico médico por meio de imagens. Contudo, eles dependem de representações de imagens que possuem alta dimensionalidade e frequentemente apresentam o problema de descontinuidade semântica.

5

Aprimoramento de rankings e de modelos de classificação via seleção de características

Neste capítulo apresentam-se as duas grandes contribuições desta tese ao campo de seleção de características visando o aprimoramento de modelos de consultas por similaridade e de classificação. As contribuições realizadas são avaliadas e discutidas no contexto de ferramentas de auxílio ao diagnóstico médico por meio de análise do conteúdo de imagens.

5.1 Considerações iniciais

No cotidiano da medicina é habitual o emprego de exames radiológicos para auxiliar no processo de diagnóstico. Este recurso é importante, mas não suficiente para a obtenção de diagnósticos corretos. A precisão de diagnóstico depende, sobretudo, de uma interpretação cuidadosa e perspicaz do caso clínico e dos exames realizados. Devido à possível falta de concentração, cansaço por longas jornadas de trabalho ou inexperiência frente a casos raros, detalhes patológicos importantes podem passar despercebidos pelos radiologistas, resultando em equívocos de diagnóstico.

Equipamentos radiológicos sem filme e os *Picture Archiving and Communication Systems* (PACSs) têm se tornado um ferramental efetivo para o arquivamento de dados clínicos. Contudo, esta valiosa fonte de conhecimento tem sido pouco aproveitada pelos médicos devido à escassez de métodos efetivos de:

- acesso e disponibilização de casos do passado (exames associados a seus diagnósticos, tratamentos e consequências) em momentos oportunos;
- previsão das classes de novos exames considerando como base os casos do passado.

O desenvolvimento de métodos efetivos de auxílio ao diagnóstico tem esbarrado nos desafios da representação do conteúdo de imagens, principalmente, na descontinuidade semântica e nos efeitos da maldição da dimensionalidade. Embora seja evidente que a seleção das características de imagem mais relevantes possam lidar com estes desafios, os métodos até então existentes têm apresentado resultados de qualidade insatisfatória para aplicações CBR e de classificação de imagens. É buscando suprir esta carência de métodos de seleção de características efetivos que se justifica os métodos apresentados nesta tese.

5.2 Introdução geral aos métodos desenvolvidos

Os métodos de seleção de características apresentados nesta tese seguem a abordagem de aprendizagem supervisionada. Dado um conjunto de exemplos rotulados na forma $(\mathbf{x}_i, s_i)$, em que $\mathbf{x}_i$ representa o vetor de características associado a uma imagem e s_i a saída desejada (Figura 5.1), deseja-se inferir um modelo ou função capaz de predizer uma saída adequada para novas imagens. Esse processo de indução de modelo, a partir de uma amostra de dados, é tradicionalmente denominado treinamento.

As saídas desejadas representam o fenômeno de interesse sobre o qual deseja-se fazer generalizações. Neste trabalho, consideram-se dois casos de valores de saída: 1) *rankings* ($\mathcal{L}$), onde cada saída s_i é uma lista ordenada de imagens, conforme a similaridade destas com a imagem de consulta $\mathbf{q}$, e busca-se selecionar as características mais adequadas para a composição de um modelo de similaridade de imagens que gera *rankings* precisos para novas imagens; e 2) rótulos (classes), onde s_i assume valores discretos $\{1, \ldots, k\}$, ou nominais, por exemplo, {"saudável", "doente"}, e busca-se encontrar as características mais adequadas para a tarefa de classificação de imagens.

Cada tupla $\mathbf{x}_i$ pertencente à tabela característica-valor $\mathbf{X}$, representada na Figura 5.1 é dada por um vetor de m valores x_{i1}, x_{i2}, x_{im} referente a um conjunto de características

previsoras $\mathbf{A} = (A_1, A_2, \ldots, A_m)$ extraído das imagens. Cada valor x_{ij} associado à característica A_j expressa um determinado aspecto (ou propriedade) de uma imagem. A ideia é que esta representação sintática capture ao máximo a semântica das imagens. O conjunto de dados tabular no formato $\langle \mathbf{x}_i, s_i \rangle$ substitui as imagens nos processos de busca e classificação.

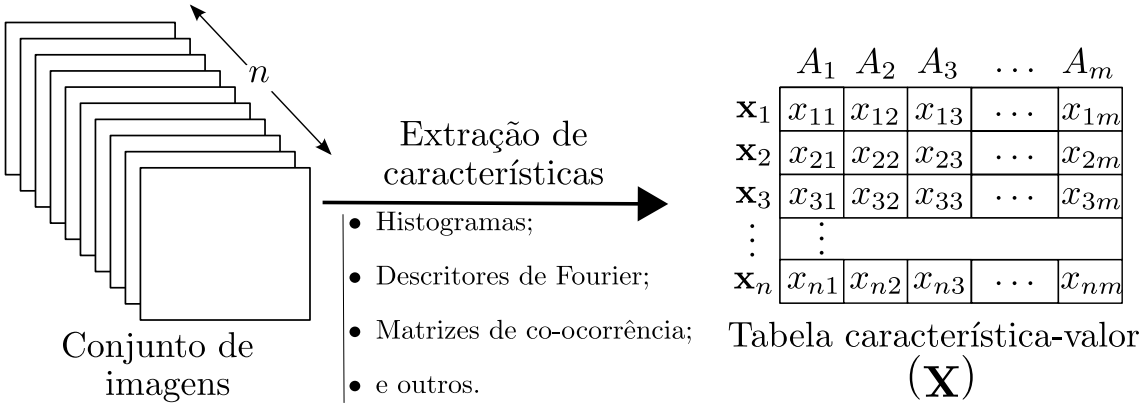


Figura 5.1: Processo de extração de características e sua representação no formato característica-valor.

Tanto os métodos de seleção de características para consulta por similaridade, quanto os para classificação de imagens propostos nesta tese, obedecem ao *pipeline* apresentado na Figura 5.2. Para permitir a validação dos resultados de seleção de características obtidos, os dados que representam as imagens no formato característica-valor são divididos em dois subconjuntos disjuntos: de treinamento e de teste. O subconjunto de treinamento é utilizado na aprendizagem do conceito meta (*rankings* apropriados ou classes) e o subconjunto de teste é utilizado para medir o grau de efetividade do conceito aprendido por meio da previsão da saída para novas imagens. Também avaliamos a taxa de redução de dimensionalidade, simbolizada como trd. Um conceito importante também discutido é o grau de generalização do modelo resultante, definido pela sua capacidade de gerar saídas corretas para novos dados. No caso em que o modelo se especializa nos dados utilizados em seu treinamento, apresentando baixa taxa de acerto quando confrontado com novos dados, tem-se a ocorrência do fenômeno clássico de super-ajustamento (*overfitting*).

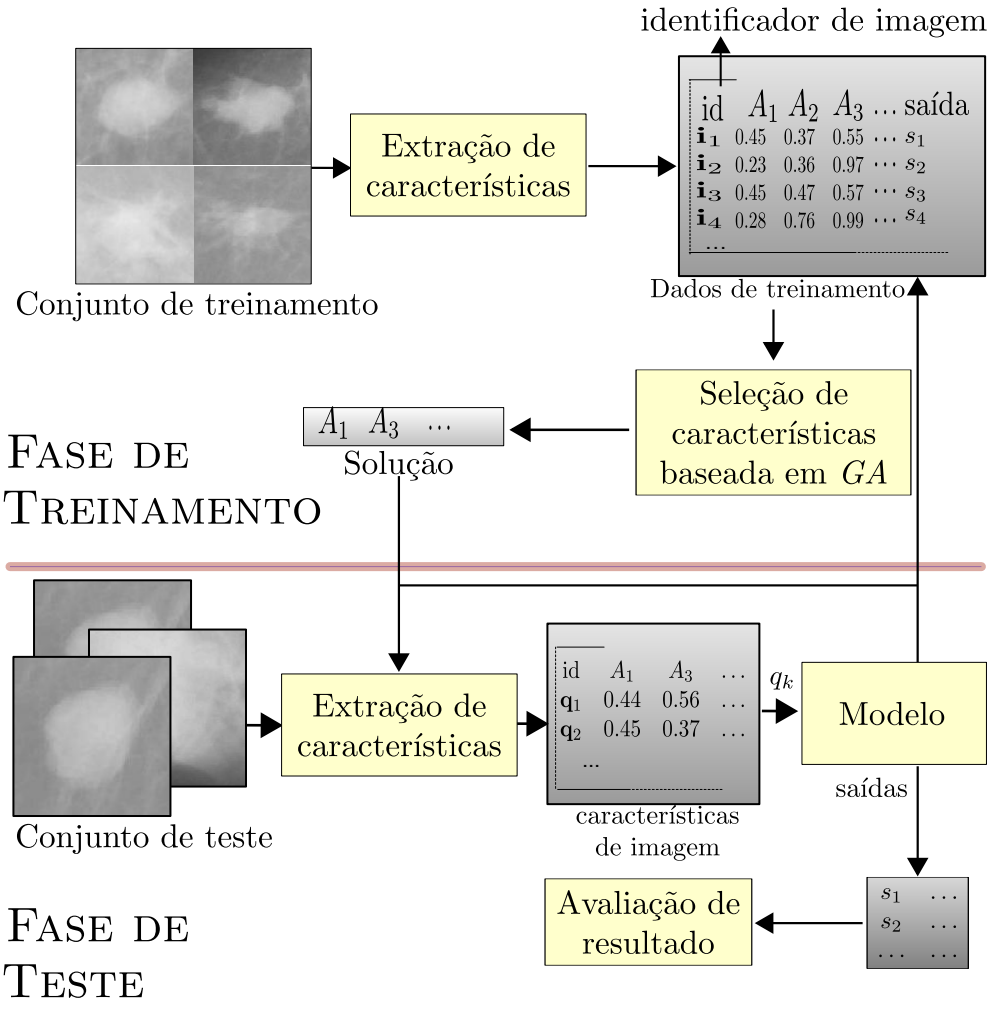


Figura 5.2: *Pipeline* geral dos métodos propostos.

5.3 Conjuntos e representações de imagens

Os conjuntos de imagens empregados nos experimentos são descritos a seguir.

5.3.1 *Mammograms ROI-250*

Conjunto de 250 imagens *ROIs* (*regions of interest*) de lesões de mama, disponibilizado pelo *Digital Database for Screening Mammography* da *University of South Carolina* em http://marathon.csee.usf.edu/Mammography/[1]. As imagens deste conjunto pertencem a duas classes: massa benigna (99 imagens) e massa maligna (151 imagens).

[1]Acessado pela última vez em 21/03/2011

5.3.2 *Mammograms-1080*

Conjunto de 1080 imagens de mamogramas realizados pelo Hospital das Clínicas de Ribeirão Preto (HCRP)-USP, classificado em 4 níveis de densidade de mama: 1) gordurosas (362 imagens), parcialmente gordurosas (446 imagens), parcialmente densas (200 imagens) e densas (72 imagens). A densidade de mama é um fator influente no desenvolvimento de câncer de mama. As imagens foram representadas com base nos extratores propostos em [58], que produz 85 características contendo informações de forma, tamanho de mama, posição do mamilo e distribuição do tecido fribroglandular.

5.3.3 *Lung ROI-3258*

Conjunto de 3258 imagens correspondentes a *ROIs* de CT (*Computed Tomography*) de pulmão, contendo seis classes, sendo uma de exame normal (590 imagens) e cinco de exames apresentando os padrões anormais: enfisema (502 imagens), consolidação (451 imagens), espessamento (590 imagens), "favo de mel" (530 imagens) e "vidro fosco" (595 imagens).

5.3.4 *ImageCLEFMed09*

Conjunto de imagens de raio-X de várias partes do corpo humano. Nos experimentos foi considerado uma amostra de 5000 imagens correspondentes às classes: crânio, espinha, braço, pulmão e perna, com 1000 imagens cada. O conjunto *ImageCLEFMed09* é disponibilizado em `http://www.imageclef.org/2009/medical/`[2].

Representações de imagens

Os conjuntos de imagens, com a exceção de *Mammograms-1080*, foram representados empregando extratores de características que capturam várias medidas das imagens considerando os aspectos de cor, forma e textura. Esta decisão foi devido ao fato de não serem conhecidas as características mais relevantes para a representação da semântica dos domínios de imagens.

[2]Acessado pela última vez em 21/03/2011.

As características de imagem extraídas de cada um dos conjuntos de imagens apresentados e as subsequentes configurações dos conjuntos de dados obtidos (dimensionalidade, número de classes, particionamento de treinamento e teste) são dadas nas Tabelas 5.1 e 5.2, respectivamente. As características extraídas foram concatenadas em um "supervetor". Como vetores de características normalmente apresentam características de diferentes magnitudes, foi empregada normalização por *z-score* (definida na Subseção 4.4.4).

	Conjuntos de imagens			
Extratores	*Lung ROI-3258*	*ImageCLEFMed09*	*Mammograms ROI-250*	*Mammograms-1080*
Momentos de Cor	-	144	-	-
Descritores de Haralick	140	88	140	-
Descritores de Sobel	-	128	-	-
Histograma de Cores	256	256	256	-
EPODHC	6	6	6	-
Tamura	6	6	6	-
Wavelet	-	64	-	-
Momentos de Zernike	255	255	255	-
Filtros de Gabor	-	48	-	-
Momentos de Hu	-	-	38	-
Run length	-	-	44	-
Extratores propostos em [58]	-	-	-	85
Dimensionalidade	707	1039	739	85

Tabela 5.1: Representação dos conjuntos de imagens empregados nas avaliações experimentais. A sigla EPODHC corresponde a estatísticas de primeira ordem derivadas do histograma de cores.

	Conjuntos de imagens			
Informações	*Lung ROI-3258*	*ImageCLEFMed09*	*Mammograms ROI-250*	*Mammograms-1080*
Número de imagens	3258	5000	250	1080
Número de classes	6	5	2	4
Instâncias de treinamento	978	1500	166	720
Instâncias de teste	2280	3500	64	360
Dimensionalidade	707	1039	739	1080

Tabela 5.2: Configuração dos conjuntos de dados empregados nos experimentos.

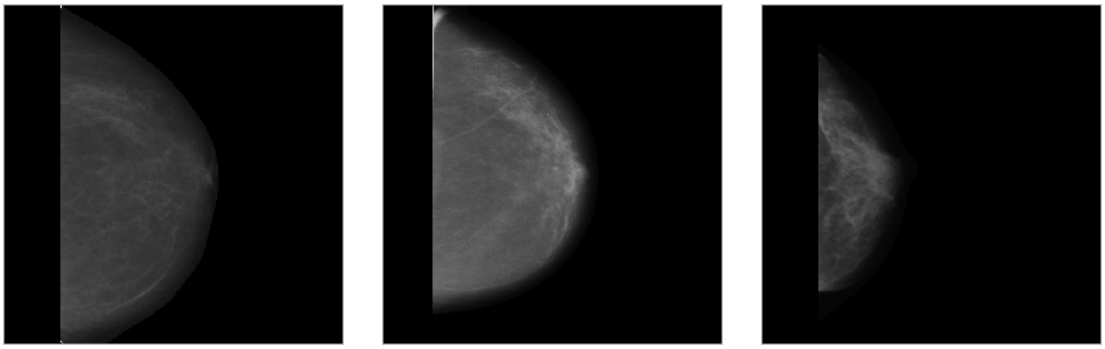
5.4 *Wrappers* de CBR

Embora recuperação de textos e de dados relacionais em geral sejam um problema resolvido pelos sistemas de recuperação de informação e de banco de dados, recuperação de imagens baseada em conteúdo permanece com vários desafios. Um dos principais é capturar e representar a semântica de similaridade em um modelo computacional.

No domínio de diagnóstico médico por imagens, o conceito de similaridade é um aspecto amplamente relacionado às patologias de interesse. Consequentemente, aspectos visuais (características) automaticamente aferidos das imagens, podem ser determinantes ou não, no estabelecimento de decisões e de relações de similaridade. Como ilustração deste fato são apresentadas na Figura 5.3 três imagens de mamografia. As imagens (a) e (b), aparentemente similares para um leigo, na realidade correspondem a patologias diferentes, ao passo que as imagens (b) e (c), que são visualmente dissimilares, correspondem a uma mesma patologia. Contudo, se for analisado particularidades específicas das imagens, tais como o aspecto de textura próximo ao mamilo, pode-se concluir que a imagem (c) é mais similar à imagem (b) do que a imagem (a). Deste modo, a escolha das características de imagens adequadas (seleção de características) é essencial para a sua análise e mensuração de similaridade do ponto vista patológico. Outros modos possíveis de adequar os mecanismos de consulta por similaridade ao domínio do problema de aplicação foram discutidos na Subseção 4.3.4 do Capítulo 4.

Embora existam várias opções de métodos de seleção de características efetivos para as áreas de classificação e agrupamento de dados, o mesmo não ocorre na comunidade CBR (que inclui a comunidade CBIR). A tarefa de seleção de características para aplicações CBR tem sido realizada por meio de métodos de filtragem e métodos *wrapper* elaborados para maximizar o desempenho de classificação, que têm, ambos, apresentado resultados insatisfatórios.

Buscando suprir a carência de métodos de seleção de características para aplicações CBR, desenvolvemos uma nova família de métodos que tem como base um conjunto de funções de avaliação de resultados de consultas por similaridade, denominado família



(a) (b) (c)

Figura 5.3: Ilustração de aspectos de similaridade patológica. As imagens (a) e (b), aparentemente similares apresentam diferentes patologias, enquanto que as imagens (b) e (c), aparentemente dissimilares em seus aspectos globais, correspondem à mesma patologia.

Fitness coach (Fc). A precisão dos resultados de seleção de características é medida em termos da corretude dos *rankings* obtidos em resposta às consultas por similaridade.

5.4.1 Definições

A fim de enquadrar os métodos propostos à taxonomia existente na literatura, foram definidas duas classes de métodos de seleção de características *wrapper*: os de classificação e os de CBR.

Definição 5.1. *Wrappers* de classificação: são métodos de seleção de características que efetuam avaliação de subconjuntos candidatos com base na acurácia do resultado de um classificador aplicado aos dados.

Definição 5.2. *Wrappers* de CBR: são métodos de seleção de características que empregam uma medida de corretude dos resultados de consultas por similaridade (*rankings*) como critério de avaliação de subconjuntos candidatos.

Os *wrappers* de classificação compõem a grande maioria dos métodos *wrappers* da literatura, conforme apresentado no Capítulo 2. *Wrappers* de CBR constituem uma nova classe de métodos *wrappers*, definidos e apresentados nesta tese. Para facilitar o entendimento dos métodos *wrappers* de CBR desenvolvidos, foram introduzidos os conceitos de *ranking*, função de avaliação de *ranking* e o critério de relevância considerado. Os conceitos de GAs necessários para este capítulo foram apresentados no Capítulo 3.

Definição 5.3. *Ranking*: Considere uma consulta aos k-vizinhos mais próximos ($kNNQ(\mathbf{q}, k, \mathbb{S})$) que recupera do conjunto $\mathbb{S}$ as k imagens mais próximos à imagem de consulta $\mathbf{q}$. A ordem das k imagens retornadas é denominada *ranking*, simbolizado por $\mathcal{L}$.

Definição 5.4. Critério de relevância: Considere um *ranking* $\mathcal{L}$ contendo k imagens obtidas em resposta a uma consulta kNN ($kNNQ(\mathbf{q}, k, \mathbb{S})$). O critério de relevância aplicado a cada imagem $\mathbf{i} \in \mathcal{L}$ é dado por uma função $r(\mathbf{i})$, onde:

$$r(\mathbf{i}) = \begin{cases} 1, & \text{se } classe(\mathbf{i}) = classe(\mathbf{q}) \\ 0, & \text{caso contrário;} \end{cases}$$

ou seja, $r(\mathbf{i})$ retorna o valor 1, quando a imagem $\mathbf{i}$ é relevante, i.e., pertence à mesma classe da imagem de consulta. Caso contrário, retorna o valor 0, indicando que a imagem $\mathbf{i}$ é irrelevante.

Definição 5.5. Função de avaliação de *ranking*: Considere um *ranking* $\mathcal{L}$ obtido como resposta a uma consulta kNN e a indicação de relevância de seus elementos dada pela função $r(\mathbf{i})$ conforme descrito na Definição 5.4. Uma função de avaliação de *ranking* $\mathfrak{F}$ provê uma nota para a corretude do *ranking* $\mathcal{L}$.

Definição 5.6. Função de avaliação de *ranking* baseada em ordem: Uma função de avaliação de *ranking* $\mathfrak{F}$ pertence à categoria baseada em ordem se ela calcula um *score* parcial para cada imagem $\mathbf{i}$ pertencente ao *ranking* $\mathcal{L}$, considerando explicitamente a posição de recuperação de $\mathbf{i}$, representada por $pos(\mathbf{i})$. Caso contrário, a função de avaliação de *ranking* é denominada não baseada em ordem.

Exemplos de funções de avaliação de *ranking* não baseada em ordem e baseadas em ordem são dados pelas Equações 4.6 e 4.7, respectivamente, apresentadas no Capítulo 4. Funções de avaliação de *ranking* baseadas em ordem baseiam-se no conceito de utilidade, onde a nota de um elemento relevante é, usualmente, inversamente proporcional à sua

posição no *ranking* (Figura 5.4). O fato de que os usuários esperam que elementos relevantes apareçam nas posições iniciais do *ranking* sugere que as funções de avaliação de *ranking* baseadas em ordem são mais bem sucedidas do que funções de avaliação não baseadas em ordem.

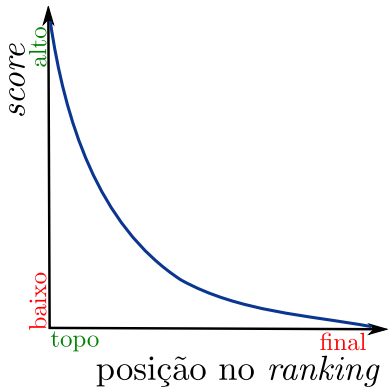


Figura 5.4: Comportamento típico dos *scores* parciais de uma função de avaliação de *ranking*, considerando a posição dos elementos no *ranking*.

Várias funções de avaliação de *ranking* têm sido propostas na literatura. Porém, conforme é conhecido, elas nunca haviam sido aplicadas para fomentar uma solução de seleção de características. Uma função de avaliação de *ranking* que tem apresentado resultados promissores para tarefas de realimentação de relevância foi apresentada na Equação 4.7 do Capítulo 4. Por ser a base para a derivação dos critérios de avaliação (família Fc) dos métodos *wrapper* de CBR, esta equação é re-apresentada abaixo e uma análise de seu parâmetro A é fornecida.

$$Fr(\mathcal{L}) = \sum_{\forall \mathbf{i} \in \mathcal{L}} \left(r(\mathbf{i}) \frac{1}{A} \left(\frac{(A-1)}{A} \right)^{(pos(\mathbf{i})-1)} \right) \tag{5.1}$$

Na Equação 5.1, $Fr(\mathcal{L})$ calcula o *score* para o *ranking* $\mathcal{L}$ considerando a função $r(\mathbf{i})$, que retorna a relevância de cada imagem $\mathbf{i}$ deste *ranking*, sendo 1 se esta for relevante e 0, caso contrário. A é um parâmetro de controle que deve assumir um valor maior ou igual a 2, podendo este ser ajustado pelo usuário. O parâmetro A indica a importância relativa da posição de elementos no *ranking*. Quando se atribuem valores baixos para A, os elementos relevantes posicionados próximo ao topo do *ranking* têm alta importância.

Quando A assume valores altos, a razão $\frac{(A-1)}{A}$ resulta em valores próximos de 1 e, assim, a posição relativa dos elementos no *ranking* não é fortemente refletida no *score* final. Na tentativa de determinar um valor equilibrado para A, onde os *scores* calculados para os elementos do *ranking* reflitam suas importâncias relativas, considerando suas relevâncias e posicionamentos e, consequentemente, fazendo com que a função $Fr(\mathcal{L})$ retrate a corretude global do *ranking* $\mathcal{L}$, foi estabelecido empiricamente, $A = 10$.

5.4.2 Família de métodos Fc

A família de métodos Fc desenvolvida é composta por uma fase de treinamento, onde as características das imagens são submetidas a um processo de seleção *wrapper* de CBR, e por uma fase de teste, onde as características selecionadas, resultantes da busca GA, são empregadas em consultas por similaridade. A curva precisão e revocação média é empregada para averiguar a eficácia dos *rankings* obtidos pelas consultas por similaridade. O *pipeline* deste processo é apresentado na Figura 5.2 e as definições e operações de GAs que implementam os métodos *wrapper* de CBR são apresentadas a seguir.

Codificação de cromossomo

A codificação de cromossomo define o modo como as soluções candidatas são representadas por meio de um arranjo de variáveis predeterminado. Nesta tese foi considerado um arranjo vetorial binário $\mathcal{C} = (g_1, g_2, \ldots, g_m)$, onde m é o número de características do conjunto de dados e cada gene g_i assume o valor 0, caso a i-ésima característica não esteja presente na solução, ou 1, caso contrário.

Operadores genéticos

GAs buscam por soluções no espaço de busca por meio das operações genéticas de seleção, cruzamento e mutação. A operação de seleção privilegia os cromossomos mais aptos, oferecendo-lhes probabilidades maiores de sobrevivência e reprodução, em relação aos menos aptos. Operações de cruzamento e mutação são analogias ao processo de reprodução natural e visam explorar o espaço de busca à procura das soluções mais eficazes. As operações genéticas empregadas neste trabalho são:

- **Seleção para cruzamento**: aplicada para selecionar pares de cromossomos para reprodução. Foi utilizada para tal, seleção por ordenação linear – os cromossomos são ordenados de acordo com seus valores de aptidão, sendo a última posição atribuída ao cromossomo mais apto. A probabilidade de seleção é distribuída linearmente conforme suas posições.

- **Seleção para reinserção**: um total de $(S_p - 2)$ cromossomos filhos e os 2 cromossomos pais, mais aptos conforme a medida de aptidão, sobrevivem da geração corrente para a próxima. S_p representa o tamanho de população empregado.

- **Cruzamento**: combina os genes de dois cromossomos (pais), gerando dois outros cromossomos (filhos). Foi empregada a operação de cruzamento uniforme, onde é preenchida aleatoriamente uma máscara binária da mesma dimensão dos cromossomos, que indica qual cromossomo pai irá fornecer cada gene para o primeiro filho. O segundo filho é gerado de maneira equivalente, utilizando uma máscara complementar à do primeiro filho.

- **Mutação**: representa a inserção de aleatoriedade no processo reprodutivo. Empregou-se a mutação uniforme onde um gene selecionado para mutação é substituído por seu complemento, isto é, transformado de 0 para 1 ou vice-versa.

Funções de aptidão

Uma função de aptidão desempenha o papel de guia da busca GA rumo às soluções mais promissoras (corretas) do espaço de busca. Funções de aptidão adequadas permitem ao GA explorar o espaço de busca de modo eficiente e eficaz, ao contrário de funções inadequadas que enfraquecem esta habilidade, podendo resultar em soluções ótimas locais. Neste trabalho, além de funções de aptidão baseadas em corretude de *ranking*, foi explorado o erro médio de classificação.

Com o intuito de construir um mecanismo geral de composição de funções de aptidão a partir de funções de avaliação de *ranking*, foi elaborado o Algoritmo 1, que efetua

validação cruzada de consultas sobre o conjunto de imagens de treinamento $\mathbf{Q}$. Na Linha 3 do Algoritmo 1, $kNNQ(\mathbf{q}, k, \mathbf{Q} - \{\mathbf{q}\}, \mathcal{C})$ corresponde a uma consulta kNN que recupera do conjunto $\mathbf{Q} - \{\mathbf{q}\}$, as k imagens mais similares a $\mathbf{q}$, considerando as características codificadas em $\mathcal{C}$. Na Linha 4, a função $\mathfrak{F}(\mathcal{L})$ corresponde a uma função de avaliação de *ranking* que estima a corretude de $\mathcal{L}$ gerado pela consulta kNN da Linha anterior. $|\mathbf{Q}|$ corresponde ao número de imagens em $\mathbf{Q}$. O processo de intercalação de consulta e base de resposta dado pelas Linhas 2 e 3 do Algoritmo 1 é similar à validação cruzada *leave-one-out* e foi elaborado para evitar *overfitting*.

Algoritmo 1: Gerador de função de aptidão a partir de consultas kNN e uma função de avaliação de *ranking* $\mathfrak{F}$.

Entrada: Conjunto de dados de treinamento $\mathbf{Q}$; cromossomo $\mathcal{C}$.
Saída: Aptidão do cromossomo $\mathcal{C}$ ($f_{\mathcal{C}}$).

1: $score = 0$;
2: **para todo** $\mathbf{q} \in \mathbf{Q}$ **faça**:
3: $\quad \mathcal{L} = kNNQ(\mathbf{q}, k, \mathbf{Q} - \{\mathbf{q}\}, \mathcal{C})$;
4: $\quad score = score + \mathfrak{F}(\mathcal{L})$;
5: $f_{\mathcal{C}} = score/|\mathbf{Q}|$
6: **retornar** $f_{\mathcal{C}}$

Com base no Algoritmo 1, foi desenvolvida a família de funções de aptidão Fc (*Fitness coach*). A aplicação do Algoritmo 1 considerando a função de avaliação de *ranking* Fr, definida na Equação 5.1, pode ser descrita como:

$$Fc(\mathbf{Q}, \mathcal{C}) = \frac{\sum_{\forall \mathbf{q} \in \mathbf{Q}} Fr(kNNQ(\mathbf{q}, k, \mathbf{Q} - \{\mathbf{q}\}, \mathcal{C}))}{|\mathbf{Q}|} \tag{5.2}$$

onde $Fr(kNNQ(\mathbf{q}, k, \mathbf{Q} - \{\mathbf{q}\}, \mathcal{C}))$ corresponde à aplicação da função de avaliação de *ranking* Fr (Equação 5.1) ao *ranking* $\mathcal{L}$ gerado pela consulta $kNNQ(\mathbf{q}, k, \mathbf{Q} - \{\mathbf{q}\}, \mathcal{C})$.

A primeira função da família Fc é então dada por:

$$FcA(\mathbf{Q}, \mathcal{C}) = 1 - \frac{Fc(\mathbf{Q}, \mathcal{C})}{\max_{\forall \mathcal{C}_j \in \mathbf{C}} Fc(\mathbf{Q}, \mathcal{C}_j)} \tag{5.3}$$

onde $\mathbf{C}$ representa a população de cromossomos.

Neste caso, o cromossomo que minimiza a Equação 5.3 representa o melhor conjunto

de características. Assim, temos um problema de minimização, que é resolvido por GA.

Vale também recordar que o princípio da seleção de características *wrapper* é minimizar o número de características selecionadas, enquanto que maximiza-se ou preserva-se a acurácia dos resultados das aplicações. Este princípio levou-nos à proposição de duas outras funções de aptidão denominadas FcB e FcC, descritas pelas Equações 5.4 e 5.5, respectivamente. Estas funções combinam dois critérios de otimização: 1) o critério de acurácia de consultas por similaridade, dado pela função FcA (Equação 5.3) e 2) a minimização do número de características selecionadas, dada por $\frac{|\Sigma\mathcal{C}-d|}{m}$ e $\frac{\Sigma\mathcal{C}}{m}$, nas Equações 5.4 e 5.5, respectivamente. Os cromossomos que resultam em valores mínimos das Equações 5.4 e 5.5 representam os subconjuntos de características ótimos procurados. Assim, temos dois novos problemas de minimização, que também foram resolvidos por GA.

$$FcB(\mathbf{Q},\mathcal{C}) = \alpha(FcA(\mathbf{Q},\mathcal{C})) + (1-\alpha)\left(\frac{|\Sigma\mathcal{C}-d|}{m}\right) \tag{5.4}$$

$$FcC(\mathbf{Q},\mathcal{C}) = \alpha(Fc(\mathbf{Q},\mathcal{C})) + (1-\alpha)\left(\frac{\Sigma\mathcal{C}}{m}\right) \tag{5.5}$$

Em ambas as Equações 5.4 e 5.5, $\Sigma\mathcal{C}$ (somatório do código binário de $\mathcal{C}$) corresponde ao número de características selecionadas, conforme o cromossomo $\mathcal{C}$, d é o número de características desejado, m é a dimensionalidade do conjunto de dados, e $\alpha \in [0,1]$ é um parâmetro que permite ao usuário ajustar a importância de cada critério de maneira complementar. A razão $\frac{|\Sigma\mathcal{C}-d|}{m}$, na Equação 5.4, produz valores altos quando o número de características selecionadas difere muito do número de características desejado pelo usuário (d). A razão $\frac{\Sigma\mathcal{C}}{m}$ é um fator de penalidade que resulta em valores máximos (próximos de 1), quando há pouca redução de dimensionalidade.

O mecanismo empregado para derivar a família de funções de aptidão Fc (FcA, FcB e FcC) se aplica a quaisquer funções de avaliação de *ranking*. Este aspecto permite experimentar funções de avaliação de *ranking* alternativas, visando a obtenção da formulação

de seleção de características mais adequada a uma dada aplicação CBR.

5.4.3 Experimentos de consultas por similaridade

Antes de serem apresentados e discutidos os resultados quantitativos obtidos, é ilustrado na Figura 5.5 o modo como um sistema de consulta por similaridade de imagens médicas provê suporte à tarefa de auxílio ao diagnóstico. Foi considerado aqui o conjunto de imagens *Mammograms ROI-250* como base de referência para as consultas por similaridade. Neste caso específico, suponha que um radiologista necessite de apoio para diagnosticar uma dada imagem recém obtida como massa benigna ou massa maligna. A imagem recém obtida é submetida ao sistema de consulta por similaridade que retornará as imagens mais similares juntamente com suas informações associadas. As informações associadas às imagens, tais como, laudos, modalidades de exames, informações do paciente, entre outras, normalmente estão contidas no cabeçalho DICOM. Com base na análise do *ranking* de imagens recuperadas e suas informações associadas, o radiologista pode encontrar rapidamente informações que o auxilie na tomada de decisão.

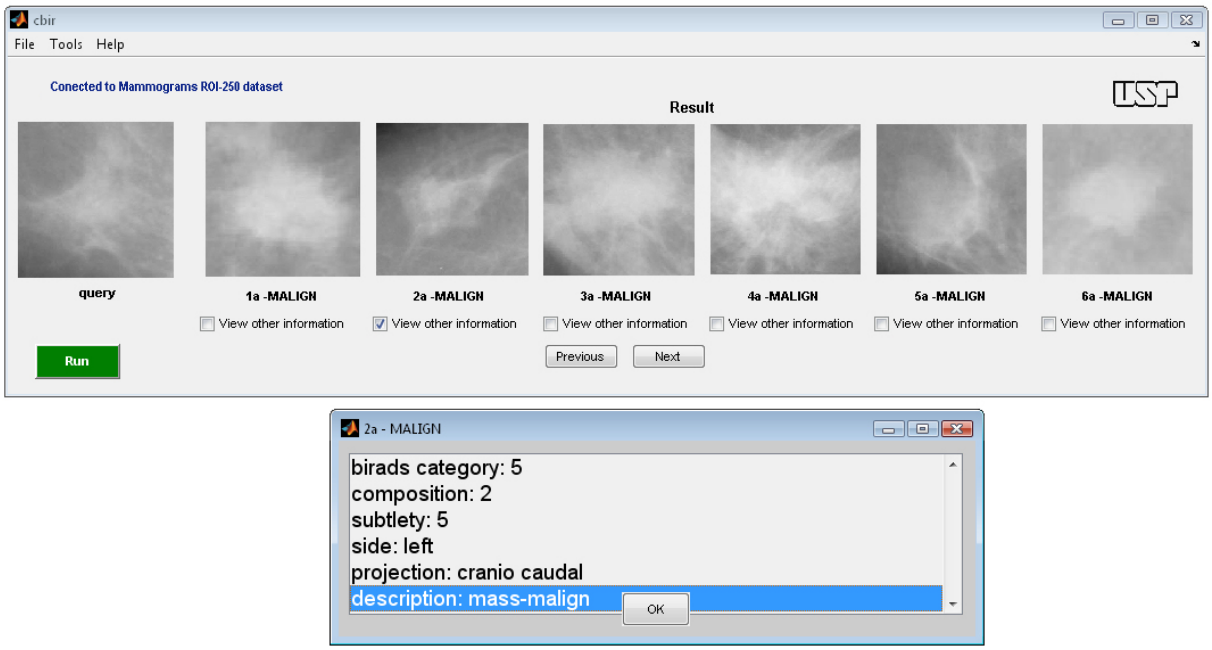


Figura 5.5: Suporte à decisão médica por meio de um resultado CBIR.

Foram realizados três experimentos onde os métodos propostos são comparados com métodos de seleção de características representativos da literatura, na tarefa de aprimo-

ramento de consultas por similaridade. Na condução dos experimentos, os conjuntos de dados foram divididos aleatoriamente em partições de treinamento e de teste. Seleções de características foram realizadas com base nos conjuntos de treinamento, sendo os resultados validados por meio de consultas tomadas dos conjuntos de teste. As avaliações de desempenho foram feitas com base na taxa de redução de dimensionalidade (trd) e por meio da análise de curvas precisão e revocação médias, considerando cada imagem do conjunto de teste como consulta e as imagens do conjunto de treinamento como base de referência para resposta.

Para facilitar a discussão e análise dos experimentos, os métodos de seleção de características foram organizados nos seguinte grupos:

(a) os métodos GA-FcA, GA-FcB, GA-FcC e GA-FR-Precision, que empregam o GA descrito na Seção 5.4.2 e as funções de aptidão FcA, FcB, FcC e FR-Precision, respectivamente. A função FR-Precision foi derivada com base no Algoritmo 1 considerando a função de avaliação de *ranking R-Precision* (Equação 4.6) e efetuando um procedimento de normalização similar da Equação 5.3. Foi também inserido neste grupo o método MS-FcA que emprega busca multipartida (*multistart search* – MS) e a função critério FcA. MS gera várias soluções iniciais aleatórias e retorna a melhor delas de acordo com a função critério aplicada. Ela tem sido empregada como base de comparação com a busca GA, visto que ambas se baseiam na geração de valores aleatórios. Espera-se que a busca GA seja significativamente superior à MS, devido a sua formulação teórica e aos mecanismos probabilísticos envolvidos. Os métodos deste grupo pertencem à categoria *wrapper* de CBR, pois eles buscam aprimorar os resultados de consultas por similaridade com base em estimativas de qualidade dos *rankings* obtidos;

(b) os métodos GA-1NN, GA-C4.5, GA-SVM e GA-NB, que empregam o mesmo GA dos métodos propostos e os erros médios de classificação de 1NN, C4.5, SVM e *Naive Bayes* (NB), respectivamente, como critério de minimização. Em cada avaliação de cromossomo (subconjunto de características), o erro médio de classificação foi obtido

por meio de validação cruzada *leave-one-out*. Os métodos deste grupo pertencem à categoria *wrapper* de classificação, pois eles visam minimizar o erro cometido por algoritmos de classificação por meio de seleção de características;

(c) o método *Statistical Association Rule Miner* (StARMiner) [91] e a não seleção de características (conjunto original);

(d) os métodos de filtragem: FCBF, *ReliefF*, CFS e mRMR;

(e) os métodos mais eficazes de cada grupo, conforme relacionado em (a)-(d).

Os experimentos com os métodos FCBF, *ReliefF* e CFS foram realizados por meio da ferramenta Weka utilizando seus parâmetros *default*. Os experimentos com mRMR foram efetuados por meio da implementação disponibilizada pelos seus autores, sendo as características previamente discretizadas pelo método *Chi2*.

A configuração do GA empregado é dada na Tabela 5.3, onde T_p é o tamanho da população empregada, P_c é a taxa de cruzamento, P_m é a probabilidade de mutação, A é o parâmetro de ajuste da função de avaliação de *ranking* da Equação 5.1, d é o número de características que deseja-se obter com a operação de seleção e α é o parâmetro de ajuste da importância de cada critério nas funções de aptidão dadas pelas Equações 5.4 e 5.5. As soluções candidatas de MS foram representadas do mesmo modo da codificação de cromossomo. Em todos os experimentos comparativos, por questão de equidade, o número de soluções aleatórias geradas por MS foi igual ao número de avaliações de aptidão realizada pelo GA.

Experimento	T_p	P_c	P_m	Gerações	A	d	α
1	100	1	0.01	400	10	50	0.9
2	50	1	0.01	250	10	20	0.9
3	100	1	0.01	400	10	100	0.9

Tabela 5.3: Parâmetros de configuração do GA empregado nos experimentos.

A Figura 5.6 mostra as curvas de precisão e revocação médias obtidas e, também, o número de características (atributos) selecionadas pelos métodos, empregando o conjunto

Mammograms ROI-250. Pode-se observar que os métodos propostos GA-FcB e GA-FcC resultaram em um aumento de precisão das consultas por similaridade, de aproximadamente 10% na faixa de até 20% de revocação, em relação ao demais métodos, enquanto que reduziram a dimensionalidade de 739 para em torno de 50 características. Mesmo empregando aproximadamente 7% das características, os métodos propostos GA-FcB e GA-FcC levaram a resultados mais precisos do que os demais métodos de seleção. A redução de dimensionalidade implica economia de espaço de memória e em redução do tempo computacional para a execução de consultas por similaridade.

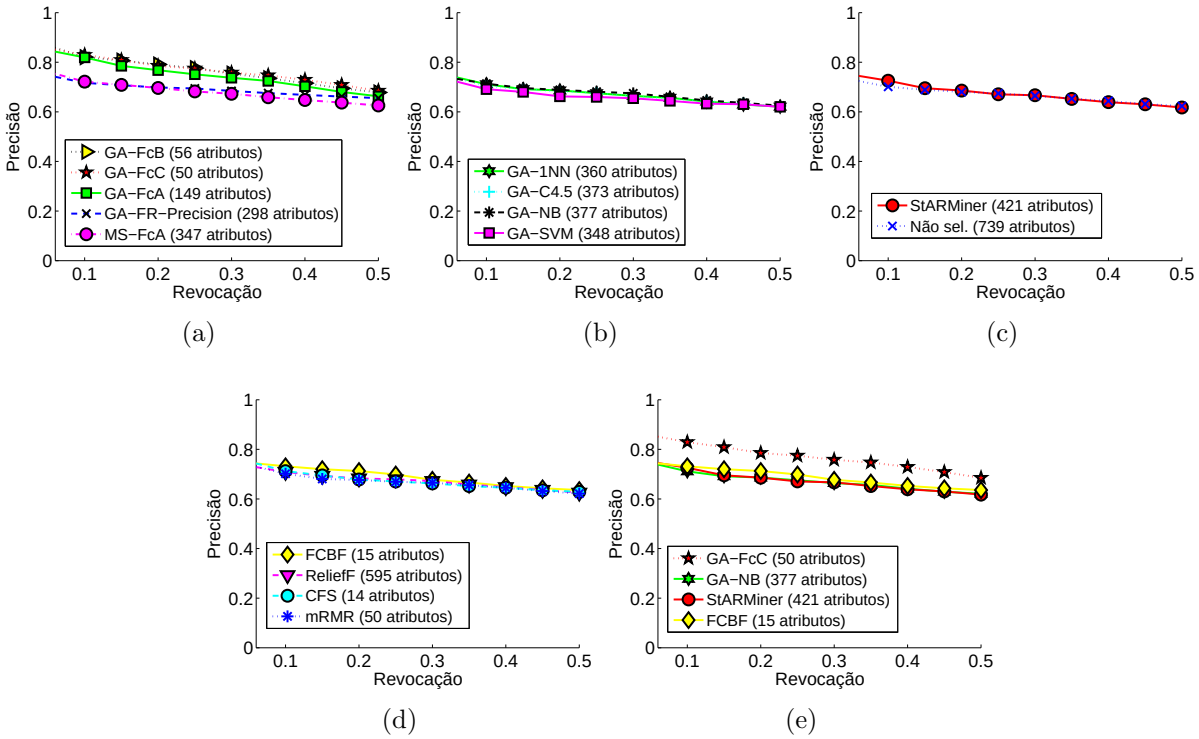


Figura 5.6: Curvas de precisão e revocação referentes ao conjunto de imagens *Mammograms ROI-250*: (a) *Wrappers* de CBR compostos por busca GA e busca multipartida (MS), empregando a família de funções critério Fc e FR-Precision; (b) *Wrappers* de classificação, consistindo de busca GA visando minimizar o erro médio cometido por classificadores tradicionais; (c) Método StARMiner e a não seleção de características; (d) Métodos de filtragem (FCBF, *ReliefF*, CFS e mRMR), e (e) os métodos mais eficazes de cada um dos grupos anteriores ((a)-(d)).

A Figura 5.7 mostra o número de características (atributos) selecionadas pelos métodos e as curvas de precisão e revocação médias, obtidas com base no conjunto *Mammography-1080*. Pode ser notado que os métodos propostos proporcionaram resultados de consultas

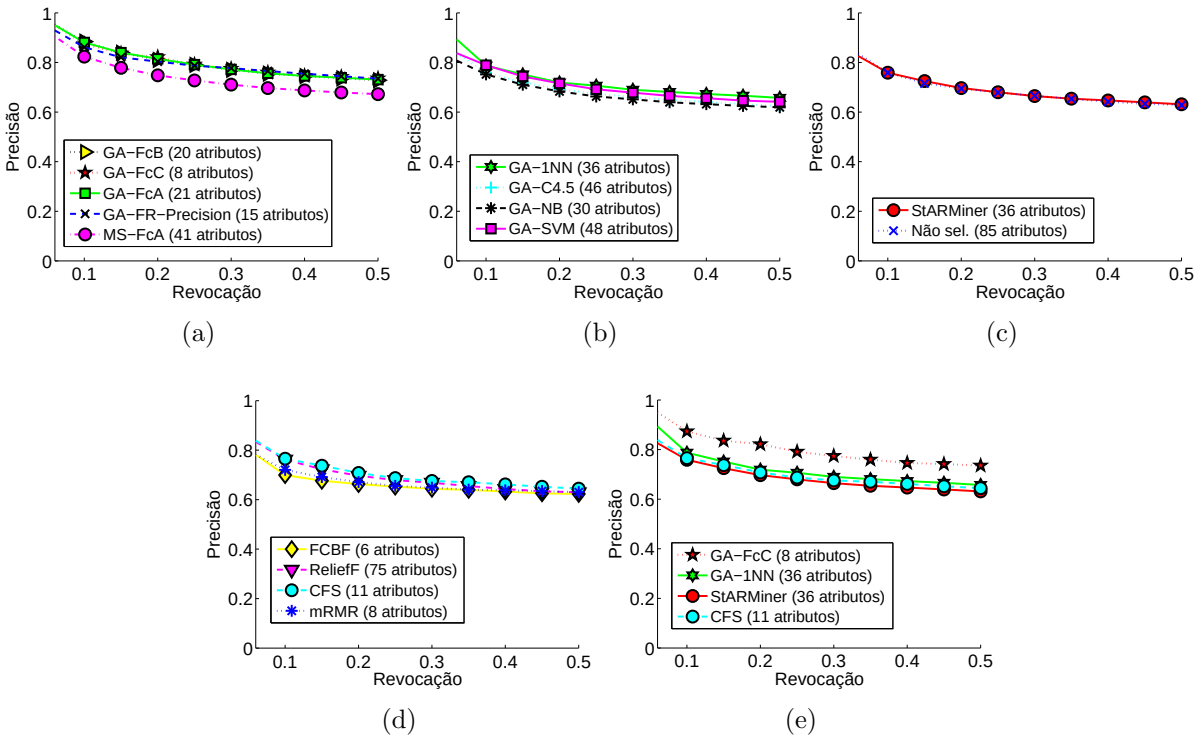


Figura 5.7: Curvas de precisão e revocação referentes ao conjunto de imagens *Mammography-1080*: (a) *Wrappers* de CBR compostos por busca GA e busca multi-partida (MS), empregando a família de funções critério Fc e FR-Precision; (b) *Wrappers* de classificação, consistindo de busca GA visando minimizar o erro médio cometido por classificadores tradicionais; (c) Método StARMiner e a não seleção de características; (d) Algoritmo de filtragem (FCBF, *ReliefF*, CFS e mRMR), e (e) os métodos mais eficazes de cada um dos grupos anteriores ((a)-(d)).

por similaridade mais precisos que os demais métodos, em aproximadamente 8% de precisão, enquanto promoveram reduções de dimensionalidade de até 90%.

A Figura 5.8 mostra as curvas de precisão e revocação obtidas, e também, o número de características selecionadas pelos métodos, quando empregados ao conjunto *Lung ROI-3258*. Pode-se observar, novamente, que a família de funções de aptidão proposta levou a um aumento significativo de precisão, selecionando aproximadamente 15% das características originais.

5.4.4 Discussão dos resultados de consultas por similaridade

Consultas por similaridade visual de imagens têm, definitivamente, grande potencialidade no domínio de auxílio ao diagnóstico médico. Contudo, sua aceitação pelos médicos

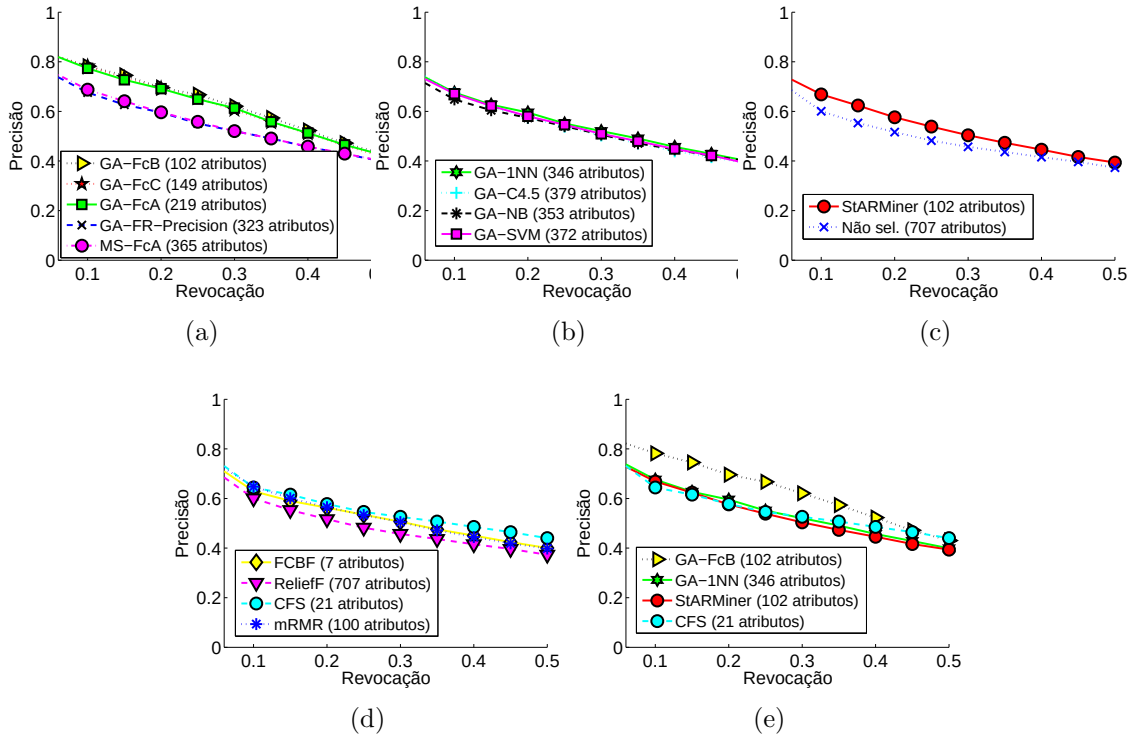


Figura 5.8: Curvas de precisão e revocação referentes ao conjunto de imagens *Lung ROI-3258*: (a) *Wrappers* de CBR compostos por busca GA e busca multipartida (MS), empregando a família de funções critério Fc e FR-Precision; (b) *Wrappers* de classificação, consistindo de busca GA visando minimizar o erro médio cometido por classificadores tradicionais; (c) Método StARMiner e a não seleção de características; (d) Algoritmo de filtragem (FCBF, *ReliefF*, CFS e mRMR), e (e) os métodos mais eficazes de cada um dos grupos anteriores ((a)-(d)).

e radiologistas dependem sobretudo de sua eficácia e eficiência. No trabalho descrito neste capítulo, procuramos atender estas demandas, introduzindo métodos de seleção de características *wrappers* especializados em otimização de *rankings* – os *wrappers* de CBR.

Foi desenvolvida uma família de métodos *wrappers* de CBR, que emprega busca GA guiada por funções de avaliação de *ranking* (família F_c), para a seleção das características mais relevantes para aplicações CBR. Os métodos de seleção de características desenvolvidos GA-FcA, GA-FcB e GA-FcC) foram comparados com: (a) métodos *wrappers* de classificação (GA-1NN, GA-C4.5, GA-SVM e GA-NB), que empregam o mesmo GA, variando apenas o módulo de avaliação de subconjuntos de características; (b) métodos de filtragem representativos da literatura (FCBF, *ReliefF*, CFS e mRMR); (c) o método StARMiner, que se baseia na mineração e análise de regras de associação e (d) a não

seleção de características (ou uso do conjunto original). Os resultados experimentais mostraram que os métodos propostos superam todos os outros métodos comparativos, provendo altas taxas de redução de dimensionalidade ao mesmo tempo que aumentam a precisão das consultas. Foi também mostrado que a busca GA supera significativamente MS em eficácia, considerando o mesmo número de cálculos de avaliação de subconjuntos candidatos. Na prática, os tempos de execução de GA-FcA e MS-FcA não diferem significativamente pois, o tempo de execução gasto é determinado pelo número de avaliações de subconjuntos de características.

Por meio da aglutinação do critério de precisão de consultas com o critério de redução da dimensionalidade, Equações 5.4 e 5.5, conseguimos reduzir significativamente a dimensionalidade dos conjuntos de dados e ainda obter ganho em precisão. Este resultado mostra que dimensionalidade baixa (em torno de, no máximo, algumas dezenas de características) é um aspecto importante para o bom desempenho de consultas por similaridade, pois, além de possíveis ganhos em precisão, o custo computacional para o processamento das consultas é reduzido significativamente.

Na Tabela 5.4 é apresentada uma comparação teórica dos métodos *wrapper* de CBR desenvolvidos com os *wrappers* de classificação e métodos de filtragem, considerando a tarefa de aprimoramento de consultas por similaridade. Os métodos híbridos não foram inclusos nesta tabela por apresentarem as mesmas limitações dos métodos *wrapper* de classificação, considerando que, até então, não existem métodos híbridos envolvendo um *wrapper* de CBR. Os métodos embutidos também não foram inseridos na tabela pois estes não podem ser diretamente aplicados para o aprimoramento de consultas por similaridade.

Os resultados apresentados neste capítulo foram publicados na forma de artigos científicos no *IEEE International Symposium on Computer-Based Medical Systems* [104] e no periódico *Decision Support Systems* (*Elsevier*) [103].

Em trabalhos futuros pretende-se: (1) aprimorar a eficiência dos métodos propostos, por meio da introdução e exploração de informações derivadas de métodos de filtragem na busca *wrapper*, além de buscar uma sinergia entre os métodos *wrapper* e de filtragem

Método	Vantagens	Limitações
***Wrappers* de classificação**	– Redução de dimensionalidade e seus benefícios consequentes	– Alto custo de avaliação de características – Risco de *overfitting* – O resultado de seleção apresenta viés em favor do classificador empregado
Filtragem	– Redução de dimensionalidade e seus benefícios consequentes – Baixo custo computacional de avaliação de características – Independente do algoritmo de aplicação	– As características selecionadas, normalmente, não são as mais relevantes para aplicações *CBR*
Wrappers* de *CBIR (**Família $\mathbf{F_c}$**)	– Redução de dimensionalidade e seus benefícios consequentes – Seleção das características mais adequadas para aplicações de consultas por similaridade	– Alto custo computacional de avaliação de características – Risco de *overfitting* – Específicos a aplicações de consultas por similaridade

Tabela 5.4: Taxonomia dos principais métodos de seleção de características aplicados no aprimoramento de consultas por similaridade. Para cada classe de métodos são apresentadas suas vantagens e limitações.

e (2) integrar informações de laudos, do histórico clínico de pacientes e de exames no mecanismo de consulta por similaridade.

5.5 Filtragem de máxima distinção

Como apresentado no Capítulo 2, um dos desafios principais da alta dimensionalidade é a indistinguibilidade de vizinhos mais próximos, que dificulta a retirada de informações e tomadas de decisões com base nos dados. Deste modo, uma operação desejável é considerar as características que resultam em máxima separabilidade dos dados.

5.5.1 Ponto de partida

Conforme discutido no Capítulo 2, os aspectos dos dados têm repercussão direta no desempenho dos métodos aplicados a estes. Do mesmo modo que determinados aspectos são maléficos, degradando o desempenho dos métodos aplicados aos dados, outros aspectos são benéficos. Na tarefa de classificação, alto grau de coesão dos objetos de uma mesma classe e alto grau de separação entre as diferentes classes são, normalmente, os aspectos mais desejados. A coesão e separação de classes de dados normalmente são mensuradas por meio de medidas de distância entre elementos, que são propriedades intrínsecas dos

dados.

Nesta parte deste trabalho buscou-se investigar a hipótese de que existe um nível de simbiose entre determinadas propriedades intrínsecas de dados e a tarefa de classificação. Para isto foi explorada a relação entre a separabilidade de classes e a acurácia de modelos de classificação. Esta relação foi avaliada por meio do desenvolvimento e experimentação de dois métodos que empregam a medida de silhueta simplificada (*simplified silhouette*), apresenta na Subseção 2.3.2 do Capítulo 2, como critério de avaliação de subconjuntos candidatos.

5.5.2 *Silhouette-based Greedy Search - SiGS*

Algoritmos de busca gulosa, tais como busca sequencial para frente, para traz e bidirecional, têm sido bastante utilizados para seleção de características, devido à sua complexidade temporal $\mathcal{O}(m^2)$, enquanto que uma busca exaustiva tem complexidade $\mathcal{O}(2^m)$, onde m é a dimensionalidade do conjunto de dados. No entanto, algoritmos de busca gulosa não provêem qualquer garantia quanto ao encontro da solução ótima, pois estes não tratam o problema de interação entre características e podem ficar presos em soluções ótimas locais. Além disso, a complexidade computacional $\mathcal{O}(m^2)$ pode ser considerada alta para seleção de características em espaços de alta dimensionalidade, tais como as representações de imagens.

Buscando amenizar o problema do tempo computacional, foi elaborado o método SiGS descrito pelo Algoritmo 2. Por este ser um procedimento guloso, ele não lida com interação de características, não garantido a obtenção da solução ótima. Contudo, ele é, na prática, mais eficiente que o algoritmo SFS tradicional, devido ao critério de parada adotado, em que a busca é encerrada no instante em que a adição de qualquer das características restantes (não selecionadas) não resultar em aumento do valor de silhueta simplificada. Como em espaços de alta dimensionalidade normalmente tem-se um número reduzido de características relevantes, esta estratégia de busca torna-se significativamente mais eficiente do que o SFS tradicional. Foi também implementado um método denominado *Silhouette-based Sequential Forward Search* (SiSFS), que emprega a busca SFS tradicional,

para servir de base de comparação com SiGS.

Algoritmo 2: *Silhouette-based Greedy Search* (SiGS).

Entrada: Conjunto de treinamento $\mathbf{Q}$, conjunto original de características $\mathbf{A}$.
Saída: Subconjunto de características selecionadas $\mathbf{A}^*$.

1: Computar o valor de silhueta simplificada $\overline{\mathfrak{ss}}$ de $\mathbf{Q}$ para cada um das características de $\mathbf{A}$. Armazenar em $\mathbf{A}^*$ a característica $A_1^* \in \mathbf{A}$ que resulta no maior valor de silhueta simplificada ($\overline{\mathfrak{ss}}$);
2: Calcular o valor de $\overline{\mathfrak{ss}}$ para todos os subconjunto formados por $\{\mathbf{A}^* \cup \{A_j\}, A_j \in \mathbf{A}, A_j \notin \mathbf{A}^*\}$. Se o valor calculado for maior que o valor $\overline{\mathfrak{ss}}$ de $\mathbf{A}^*$, então fazer $\mathbf{A}^* = \mathbf{A}^* \cup \{A_j\}$; Repetir o Passo 2;
3: Retornar $\mathbf{A}^*$;

5.5.3 *Silhouette-based Genetic Algorithm Search - SiGAS*

Como apresentado no Capítulo 2, os métodos de busca sequencial tais como SFS, SBS e os algoritmos gulosos em geral não lidam com o aspecto de interação entre características dos conjuntos de dados, permitindo a eliminação de características que, em conjunto com outras, são altamente relevantes. Visando tratar este aspecto foi proposto o método SiGAS, que conta com um mecanismo de busca baseado em GA. A propriedade de busca global com componentes aleatórios e probabilísticos dos GA*s* permite lidar naturalmente com a interação entre características.

O método SiGAS é apresentado por meio da descrição dos passos do GA que o implementa. Um dos requisitos para resolução de problemas por *GA* é a definição da representação de cromossomo. Neste desenvolvimento, os cromossomos foram representados por um vetor binário $\mathcal{C} = (g_1, g_2, \ldots, g_m)$, onde m é a dimensionalidade do conjunto de dados e cada gene g_i assume o valor 0, caso a i-ésima característica não seja selecionada, ou o valor 1, caso contrário. Os cromossomos (representações de subconjuntos de características selecionadas) foram avaliados por meio da medida de silhueta simplificada considerando um conjunto de treinamento $\mathbf{Q}$. O método de seleção por ordenação linear foi aplicado para a seleção de genitores para reprodução. No processo de reprodução foram empregados os operadores de cruzamento uniforme e de mutação uniforme. Os parâmetros empregados na busca *GA* são dados na Tabela 5.5, onde T_p é o tamanho da

população empregada, P_c é a taxa de cruzamento, P_m é a probabilidade de mutação (por gene) e elitismo corresponde à quantidade de cromossomos elite da geração corrente que são mantidos na próxima.

Conjunto de dados	T_p	P_c	P_m	Gerações	Elitismo
Lung ROI-3258	100	1	0.03	350	3
ImageCLEFMed09	200	1	0.03	500	3
Mammograms ROI-250	100	1	0.03	350	3

Tabela 5.5: Parâmetros de configuração do GA empregado nos experimentos.

Algoritmo 3: *Silhouette-based Genetic Algorithm Search* (SiGAS).

Entrada: Conjunto de treinamento **Q**, conjunto original de características **A**.
Saída: Conjunto de características selecionadas $\mathbf{A}^*$.

1: Gerar uma população aleatória de cromossomos;
2: Avaliar a aptidão de cada cromossomo por meio da medida de silhueta simplificada;
3: Aplicar o operador de seleção por ordenação linear para selecionar os pares de cromossomos para reprodução;
4: Aplicar os operadores de cruzamento e mutação;
5: Substituir os cromossomos da geração anterior pelos gerados no Passo 4, considerando elitismo de três cromossomos;
6: Enquanto o número máximo de gerações não é atingido, retornar ao Passo 2;
7: Retornar o subconjunto de características $\mathbf{A}^* \subseteq \mathbf{A}$, dado pelo cromossomo mais apto.

5.5.4 Experimentos de classificação

A experimentação dos métodos desenvolvidos foi realizada por meio de classificação baseada em conteúdo empregando os conjuntos de imagens e dados *Lung ROI-3258*, *ImageCLEFMed09* e *Mammograms ROI-250* descritos na Seção 5.3.

Os métodos desenvolvidos foram comparados com o método *wrapper K-Nearest Neighbor-based Genetic Algorithm Search* (kNNGAS) e com os métodos de filtragem SiSFS, CFS, FCBF e *ReliefF* com base na taxa de classificação correta (acurácia) dos classificadores kNN ($k = 1$) e *Naive Bayes* (NB) e na taxa de redução de dimensionalidade trd. O método kNNGAS emprega a mesma busca GA de SiGAS buscando maximizar a acurácia de classificação do método kNN. Os métodos CFS, FCBF e *ReliefF* (descritos no Capítulo 2) e os classificadores kNN ($k = 1$) e NB (descritos no Capítulo 4) foram

executados por meio da ferramenta Weka utilizando os seus parâmetros *default*.

	Conjuntos de imagens								
	Lung ROI-3258 ($m = 707$)			*ImageCLEFMed09* ($m = 1039$)			*Mammograms ROI-250* ($m = 739$)		
Método	%train.	%teste	trd	%train.	%teste	trd	%train.	%teste	trd
Não sel./kNN	76.84^{d}	72.65^{c}	0.00^{e}	54.61^{i}	55.13^{i}	0.00^{f}	69.94^{gh}	66.57^{d}	0.00^{e}
SiSFS/kNN	79.30^{c}	75.84^{b}	98.56^{b}	79.76^{c}	78.96^{b}	93.66^{c}	65.92^{i}	66.31^{d}	99.69^{a}
SiGS/kNN	78.89^{c}	75.55^{b}	98.61^{ab}	78.36^{d}	77.19^{c}	96.70^{b}	65.92^{i}	66.3^{d}	99.69^{a}
SiGAS/kNN	82.29^{b}	78.51^{a}	98.39^{b}	83.93^{ab}	83.00^{a}	92.96^{c}	68.62^{h}	67.50^{cd}	99.36^{a}
kNNGAS/kNN	85.25^{a}	77.06^{ab}	50.51^{d}	84.13^{a}	78.82^{b}	70.90^{d}	73.39^{ef}	65.60^{d}	54.84^{c}
CFS/kNN	78.99^{c}	75.31^{b}	96.32^{c}	82.94^{b}	82.14^{a}	96.10^{b}	72.93^{ef}	66.70^{d}	97.48^{b}
FCBF/kNN	71.98^{e}	69.70^{d}	99.12^{a}	75.59^{e}	75.08^{c}	98.39^{a}	74.77^{def}	65.13^{d}	97.79^{b}
ReliefF/kNN	76.86^{d}	73.49^{c}	0.13^{e}	73.19^{f}	71.97^{e}	49.17^{e}	71.95^{fg}	66.02^{d}	15.37^{d}
Não sel./NB	62.69^{hi}	62.77^{e}	0.00^{e}	48.19^{k}	48.12^{k}	0.00^{f}	73.85^{def}	72.49^{ab}	0.00^{e}
SiSFS/NB	68.70^{fg}	68.57^{d}	98.56^{b}	64.46^{g}	64.35^{g}	93.66^{c}	76.26^{bc}	75.32^{ab}	99.69^{a}
SiGS/NB	68.50^{g}	68.19^{d}	98.61^{ab}	64.51^{g}	64.15^{g}	96.70^{b}	76.26^{bc}	75.32^{ab}	99.69^{a}
SiGAS/NB	69.88^{f}	69.66^{d}	98.39^{b}	65.59^{g}	66.90^{f}	92.96^{c}	77.06^{b}	77.15^{a}	99.36^{a}
kNNGAS/NB	63.61^{h}	63.75^{e}	50.51^{d}	56.40^{h}	57.27^{h}	70.90^{d}	86.48^{a}	72.36^{b}	54.84^{c}
CFS/NB	63.32^{h}	63.36^{e}	96.32^{c}	64.86^{g}	64.82^{g}	96.10^{b}	77.40^{b}	74.40^{ab}	97.48^{b}
FCBF/NB	61.74^{i}	62.25^{e}	99.12^{a}	65.24^{g}	64.92^{g}	98.39^{a}	75.92^{bcd}	72.24^{b}	97.79^{b}
ReliefF/NB	62.71^{hi}	62.75^{e}	0.13^{e}	52.03^{j}	52.02^{j}	49.17^{e}	73.57^{ef}	72.11^{bc}	15.37^{d}
CV	2.15	2.85	0.26	1.97	1.94	0.54	3.42	1.03	3.78
LSD	1.37	1.76	0.53	1.19	1.16	1.25	2.23	4.72	0.98

Tabela 5.6: Desempenho dos métodos de seleção de características analisados, empregando *Least Significant Difference* (LSD) *t-test* com probabilidade $p = 0.05$. Os valores de coeficiente de variação (CV) e *LSD* nas duas últimas linhas da tabela correspondem ao coeficiente de variação e diferença mínima significativa do teste, respectivamente.

Cada medida de acurácia de classificação é, na realidade, correspondente ao arranjo: método de seleção/classificador, indicado na primeira coluna da Tabela 5.6. Na Tabela 5.6, o termo "Não sel." significa que no presente experimento foram consideradas todas as características do conjunto de dados, i.e., não foi realizada seleção de características; "%train." e "%teste" correspondem às porcentagens de acurácia obtidas no treinamento e teste, respectivamente, e trd denota a taxa de redução de dimensionalidade. Os valores de acurácia e de redução de dimensionalidade foram obtidos por meio de 10 execuções em cada experimento, considerando 10 partições aleatórias de cada conjunto de dados em treinamento e teste.

A acurácia de cada execução de classificação (nos conjuntos de treinamento e de teste)

foi obtida sob validação cruzada *k-fold*, com $k = 10$. O teste estatístico LSD *t-test* com probabilidade $p = 5\%$ (confiança de 95%) foi empregado para respaldar a retirada de conclusões, com um suporte estatístico. Esta concepção de experimentação é necessária pois diferentes particionamentos do conjunto de dados resultam em diferentes resultados de seleção de características e, consequentemente, de classificação. Resumidamente, o LSD *t-test* calcula a média obtida pelas r repetições de um experimento e então, os valores do coeficiente de variação (CV) e da diferença mínima significativa (LSD). Com base nas médias obtidas e no valor de LSD, os resultados de experimentos concorrentes são sintetizados por meio de letras. Médias seguidas pela mesma letra indicam que os testes não diferem estatisticamente conforme a probabilidade p empregada e a ordem das letras fornece o *ranking* dos testes, sendo 'a' superior a 'b', e assim sucessivamente.

Resultados de seleção de características para classificação normalmente são analisados com base na taxa de redução de dimensionalidade (trd) e na acurácia de classificação nas fases de treinamento e de teste. A redução de dimensionalidade é fundamental para aumentar a eficiência computacional e diminuir os riscos de *overfitting* de modelos de classificação, além de auxiliar na compreensibilidade dos modelos e conjuntos de dados, e resultar em resultados mais precisos. Enquanto que acurácia máxima de classificação em teste é um fator primordial em seleção de características, acurácia máxima em treinamento não é importante. Contudo, a análise pareada destas duas medidas permite verificar a capacidade de generalização de um dado modelo de classificação. Analisando a Tabela 5.6 podemos notar que:

- a acurácia do arranjo kNNGAS/kNN foi alta na fase de treinamento para todos os conjuntos de dados, porém sua superioridade não se manteve na fase de teste, indicando um leve *overfitting* deste arranjo;

- considerando a acurácia na fase de teste como objetivo primordial e a redução de dimensionalidade como critério de desempate, pode-se verificar que o método proposto SiGAS obteve resultados superiores aos demais;

- o método SiGAS obteve acurácia de classificação significativamente superior aos métodos SiGS e SiSFS. Este resultado ilustra o fato dos GAs lidarem com a interação entre características de forma mais eficiente do que SFS e GS. Este resultado também indica a existência de interação entre características em representações de imagens;

- os resultados dos métodos SiGS e SiSFS, na maioria das vezes, não diferiram estatisticamente. Contudo, *SiGS* é a melhor opção quando consideramos o tempo real de processamento, devido a sua heurística de encerramento da busca. O número de cálculos de silhueta economizado por SiGS em relação a SiSFS é proporcional a taxa de redução de dimensionalidade;

- o classificador kNN obteve desempenho superior ao NB nos conjuntos de dados *Lung ROI-3258* e *ImageCLEFMed09* que têm, ambos, alguns milhares de elementos, porém, a situação se inverteu no conjunto *Mammograms ROI-250*, que tem apenas 250 elementos. Este resultado é devido a propriedades bem conhecidas dos classificadores kNN e NB em função da configuração dos conjuntos de dados. Como conhecido da literatura, o classificador kNN normalmente apresenta baixo desempenho quando o conjunto de dados tem alta dimensionalidade e poucas instâncias, e seu desempenho tende a aumentar à medida que esta situação se inverte. Já o classificador NB, por ser um método estatístico, é menos vulnerável do que o kNN a conjuntos de dados esparsos, como é o caso do conjunto *Mammograms ROI-250*.

- seleção de características empregando os métodos propostos resultou em redução de dimensionalidade acima de 92% em todos os conjuntos de dados com ganhos de acurácia na fase teste de até 28%.

5.5.5 Discussão dos resultados de classificação

Classificação é uma das tarefas computacionais mais utilizadas atualmente no campo de apoio ao diagnóstico. Contudo sua ampla aceitação e difusão têm esbarrado em sua

baixa eficácia para muitas aplicações de imagens. Este problema tem sido tratado mas não satisfatoriamente resolvido por técnicas de seleção de características *wrappers* e de filtragem. Métodos de seleção de características *wrapper* apresentam alto custo computacional e têm se mostrado propensos a *overfitting* em tarefas envolvendo dados de alta dimensionalidade, bastante comuns em classificação de imagens. Métodos de filtragem são, em geral, menos propensos a *overfitting*, contudo, em muitos casos as características selecionadas não são as mais relevantes para a tarefa de classificação.

Diante do impasse de custo-benefício dos métodos *wrappers* e de filtragem, foi lançada e trabalhada a hipótese de que existem simbioses entre propriedades intrínsecas dos dados e a tarefa de classificação. Considerando esta hipótese foram desenvolvidos dois métodos de filtragem que objetivam maximizar a separabilidade entre as classes do conjunto de dados: SiGS e SiGAS, sendo que ambos se baseiam no critério de silhueta que permite identificar o conjunto de características que provê a separabilidade máxima entre as classes. Este aspecto se mostrou bastante promissor na seleção das características mais relevantes para algoritmos de classificação, superando métodos de filtragem tradicionais (CFS, FCBF e *ReliefF*) e o método *wrapper* kNNGAS que é um dos mais eficazes da literatura.

Os experimentos, ilustrados pela Tabela 5.6, mostraram que os métodos propostos confirmam a hipótese lançada sobre a existência de propriedades intrínsecas de alta simbiose com a tarefa de classificação. Quando a medida de propriedade intrínseca dos dados é escolhida adequadamente, os métodos de filtragem competem com métodos *wrapper* de classificação em termos de acurácia dos resultados proporcionados. Adicionalmente, os métodos de filtragem apresentam as vantagens de demandarem menor custo computacional, de apresentarem propensão mínima a *overfitting* e de serem independentes do método de classificação utilizado na etapa posterior.

Na Tabela 5.7 é apresentada uma comparação teórica das principais classes de métodos de seleção de características aplicados no aprimoramento de modelos de classificação. Os métodos híbridos não foram inclusos nesta tabela pois eles apresentam, basicamente, as mesmas características dos métodos *wrapper* de classificação, com um custo computacional

ligeiramente reduzido, devido a redução do número de avaliações do tipo *wrapper*.

Método	Vantagens	Limitações
***Wrappers* de classificação**	– Seleciona, conforme o conjunto de treinamento, as características mais adequadas para um dado algoritmo de classificação	– Alto custo de avaliação de características – Risco de *overfitting* – O resultado de seleção apresenta viés em favor do classificador empregado
Filtragem da literatura	– Baixo custo computacional de avaliação de características – Independente do algoritmo de aplicação	– As características selecionadas, normalmente, não são as mais relevantes para a tarefa de classificação
Embutidos	– Custo computacional intermediário – Selecionam as características mais relevantes ao mesmo tempo que gera o classificador	– Seleção de características dependente do método de inferência empregado – Risco de *overfitting*
Filtragem de silhueta (propostos)	– Baixo custo computacional de avaliação de características – Seleciona as características mais adequadas para vários métodos de classificação	– Aplicável somente a espaços métricos – Resultado de seleção apresenta viés em favor de tarefas de classificação

Tabela 5.7: Taxonomia dos principais métodos de seleção de características aplicados no aprimoramento de modelos de classificação. Para cada classe de métodos são apresentadas suas vantagens e limitações.

Os resultados apresentados nesta seção foram publicados na forma de artigo científico no *IEEE International Symposium on Computer-Based Medical Systems* [102].

5.6 Considerações finais

Neste capítulo apresentaram-se as contribuições de seleção de características para as tarefas de consulta por similaridade e classificação de imagens. Os experimentos apresentados para cada uma delas (Seções 5.4 e 5.5, respectivamente), mostram que os métodos propostos são superiores aos concorrentes na literatura. As Subseções 5.4.4 e 5.5.5 discutem, em detalhes, os resultados obtidos para consulta de similaridade e classificação.

As conclusões e trabalhos futuros relacionados à esta tese são dados no próximo capítulo.

6 Conclusões e trabalhos futuros

6.1 Considerações iniciais

Métodos efetivos de consulta por similaridade e classificação de imagens são altamente almejados no desenvolvimento de sistemas de apoio ao diagnóstico médico. Contudo, a elaboração de tais métodos tem esbarrado nos desafios encontrados na representação do conteúdo de imagens, que são a descontinuidade semântica e a maldição da dimensionalidade. Deste modo, a redução da descontinuidade semântica e mitigação dos efeitos da maldição da dimensionalidade constituem desafios que ainda carecem de pesquisa e desenvolvimento. Uma das técnicas que tem grande potencialidade neste domínio é a seleção das características mais relevantes para as tarefas de classificação e recuperação por conteúdo.

Vários métodos de seleção de características têm sido propostos ao longo das últimas décadas. Todavia, as pesquisas abordavam o problema de seleção de características para aplicações CBR por meio de métodos não especializados à tarefa, que consequentemente, não realizavam um aprimoramento efetivo das consultas por similaridade. A seleção das características mais adequadas para responder consultas por similaridade permite contornar os efeitos da maldição da dimensionalidade ao mesmo tempo em que reduz a descontinuidade semântica, por considerar as características mais importantes no estabelecimento de relações de similaridade conforme a semântica do domínio da aplicação.

Outro ponto importante que não havia sido investigado na literatura sobre seleção

de características é a possibilidade de definir métodos de filtragem de alta simbiose com determinadas tarefas. Este tópico de investigação é importante pois os algoritmos de filtragem normalmente apresentam baixo custo computacional e são minimamente suscetíveis a *overfitting*.

6.2 Principais contribuições

Esta tese contribui ao avanço científico no que concerne aos métodos de seleção de características, endereçando principalmente a sua aplicação no aprimoramento de CBR (que executam consultas por similaridade) e de classificação de dados (que servem como "segunda opinião" no apoio ao diagnóstico médico), em situações de alta dimensionalidade. As principais contribuições desta tese são:

- definição da abordagem *wrapper* de CBR – uma nova classe de métodos de seleção de características dedicada ao aprimoramento de consultas por similaridade;
- desenvolvimento de uma família de funções de avaliação de características (família "*Fitness coach*" (Fc)) apoiando-se em conceitos de qualidade de *ranking*;
- desenvolvimento de métodos de seleção de características via busca GA, guiada pelas funções de avaliação da família Fc;
- definição e confirmação da hipótese de que existe um nível de simbiose significativo entre determinadas propriedades intrínsecas de conjunto de dados e o desempenho de métodos de classificação, mais especificamente:
  - que alta separabilidade entre classes é um aspecto importante para o desempenho de métodos de classificação e, consequentemente;
  - que a busca pelas dimensões dos dados que resultam no maior valor de separabilidade entre classes conforme a medida de silhueta, considerando a sua versão simplificada, permite encontrar as características que levam ao melhor desempenho dos classificadores kNN e NB;

- experimentação mostrando a supremacia dos métodos *wrapper* de CBR em relação aos métodos *wrapper* de classificação e de filtragem, na seleção das características mais relevantes para responder consultas por similaridade;

- resultados conclusivos de que os GAs levam a resultados significativamente superiores aos métodos de busca sequencial. Este acontecimento pode ser explicado pelo fato dos GAs lidarem naturalmente com interação entre características, além de raramente ficarem presos em soluções mínimas locais;

Outra contribuição relacionada a esta tese, consistiu do emprego de funções de avaliação de *ranking* para a otimização de CBIR por meio de pesos de características. Foram analisadas dez funções de avaliação de *ranking* que seguem duas abordagens: baseadas em ordem e não baseadas em ordem. Desta análise concluiu-se que funções de avaliação de *ranking* baseadas em ordem superam as não baseadas em ordem, uma vez que é obtido um número maior de imagens relevantes próximas ao topo to *ranking*, além de tornar a busca GA mais eficiente. Foi elaborado um artigo científico contendo os resultados desta pesquisa, o qual foi submetido ao periódico *Pattern Recognition Letters (Elsevier)*.

6.3 Trabalhos futuros

As contribuições apresentadas nesta tese geraram a necessidade de novos estudos, tanto para estender as técnicas desenvolvidas e experimentar novas formulações de critérios de seleção de características, quanto para abordar outros fatores.

Quanto a abordagem *wrapper* de CBR pretende-se:

- analisar uma gama maior de funções de avaliação de *rankings*, conforme foi feito no artigo submetido ao periódico *Pattern Recognition Letters*.

- comparar os *wrappers* de CBR com métodos de aprendizagem de funções de similaridade entre imagens;

- identificar associações entre as característica selecionadas e os parâmetros perceptuais empregados pelos médicos na identificação de patologias e no estabelecimento

de relações de similaridade;

- aprimorar a eficiência dos métodos por meio da inserção de busca local à busca global efetuada pelo GA.

A respeito dos métodos de filtragem de máxima distinção cogita-se:

- aprimorar o índice de silhueta, modificando a medida de dissimilaridade entre as instâncias e os *clusters* (grupos) do conjunto de dados;
- estudar e avaliar outros de índices de separabilidade entre classes, tais como: Jaccard, Davies-Bouldin e Calinski-Harabasz;
- investigar a combinação de índices de separabilidade entre classes.

6.4 Publicações

Considera-se a produção de artigos científicos uma forma de se validar a pesquisa desenvolvida em um projeto de doutorado, como a que culminou nesta tese. Os artigos principais publicados durante o doutoramento, em periódicos e conferências internacionais, além dos submetidos são apresentados na tabela 6.4.

Ano	Título	Conferência/ Periódico	Contribuição
2007	*Adaptive Image Retrieval through the use of a Genetic Algorithm [100]*	*IEEE International Conference on Tools with Artificial Intelligence*	Mecanismo de ponderação de características de imagens baseado em GA. Foi desenvolvida uma função critério que considera o *feedback* do usuário e as posições das imagens no *ranking* resposta das consultas.
2009	*Ranking Evaluation Functions to Improve Genetic Feature Selection in Content-Based Image Retrieval of Mammograms [104]*	*IEEE International Symposium on Computer-Based Medical Systems (CBMS)*	Descrição dos métodos da família de funções critério Fc e experimentos iniciais.
2010	*Silhouette-based feature selection for classification of medical images [102]*	*IEEE International Symposium on Computer-Based Medical Systems*	Métodos de filtragem de máxima distinção, empregando o índice silhueta simplificada como função critério.
2010 (Submetido)	*RaCBIR: a content-based image retrieval system based on ranking optimization*	*Pattern Recognition Letters - Elsevier*	Extensão do artigo [100]. São experimentadas oito novas funções de avaliação de *ranking*; realizada a análise de gráficos de P&R; e estudada a complexidade do algoritmo.
2011	*Improving the ranking quality of medical image retrieval using a genetic feature selection method [103]*	*Decision Support Systems - Elsevier*	Extensão do artigo [104]. Descrição detalhada dos métodos da família Fc, experimentação envolvendo novos conjuntos de dados e comparação da busca GA com *Multistart Search* (MS).
2011 (Submetido)	*H-Metric: characterizing image datasets via homogenization based on kNN-queries*	*Data Science Journal*	Descrição e avaliação de *H-Metric*, que é uma métrica de estimação da complexidade semântica de conjuntos de imagens com base em modificações controladas em suas classes.

Tabela 6.1: Principais artigos produzidos durante o período de doutorado.

Referências Bibliográficas

[1] Aine, S., Kumar, R., and Chakrabarti, P. (2009). Adaptive parameter control of evolutionary algorithms to improve quality-time trade-off. *Applied Soft Computing*, 9:527–540. (Citado na página 39.)

[2] Al-Kadi, O. S. (2010). Assessment of texture measures susceptibility to noise in conventional and contrast enhanced computed tomography lung tumour images. *Computerized Medical Imaging and Graphics*, 34:494–503. (Citado nas páginas 2 e 44.)

[3] Alfaro-Cid, E., McGookin, E., and Murray-Smith, D. (2009). A comparative study of genetic operators for controller parameter optimisation. *Control Engineering Practice*, 17:185–197. (Citado na página 39.)

[4] Alto, H., Rangayyan, R. M., and Desautels, J. E. L. (2005). Content-based retrieval and analysis of mammographic masses. *Journal of Electronic Imaging*, 14(2):1–17. (Citado na página 45.)

[5] Antani, S., Long, L. R., and Thoma, G. R. (2008). Bridging the gap: Enabling cbir in medical applications. In *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 4–6. (Citado nas páginas 2 e 42.)

[6] Arbib, M. A. (2003). *The Handbook of Brain Theory and Neural Networks*. MIT Press, Massachusetts, England, 2nd edition. (Citado na página 26.)

[7] Arivazhagan, S. and L., G. (2003). Texture classification using wavelet transform. *Pattern Recognition Letters*, 24:1513–1521. (Citado na página 44.)

[8] Austin, S. (1990). *An Introduction to Genetic Algorithms*. AI Expert. (Citado na página 28.)

[9] Bäck, T., Fogel, D. B., and Michalewicz, Z. (2000). *Evolutionary Computation 1: Basic Algorithms and Operators*, volume 1. Institute of Physics Publishing, Philadelphia, USA. (Citado na página 26.)

[10] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley, Essex, UK. (Citado nas páginas 49 e 50.)

[11] Balan, A. G. R. (2007). *Métodos adaptativos de segmentação aplicados à recuperação de imagens por conteúdo*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP. (Citado na página 46.)

[12] Bartell, B., Cottrell, G., and Belew, R. (1998). Optimizing similarity using multi-query relevance. *Journal of the American Society for Information Science*, 49:742–761. (Citado na página 49.)

[13] Bellman, R. (1961). *Adaptive control processes: a guided tour*. Princeton University Press. (Citado nas páginas 2 e 11.)

[14] Bermejo, P., de la Ossa, L., Gámez, J. A., and Puerta, J. M. (2011). Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems*. (to appear). (Citado na página 18.)

[15] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbour" meaningful? In *Proceedings of the 7th International Conference on Data Theory, LNCS, Springer-Verlag*, volume 1540, pages 217–235. (Citado nas páginas 2 e 11.)

[16] Bianconi, F. and Fernández, A. (2007). Evaluation of the effects of gabor filter parameters on texture classification. *Pattern Recognition*, 40:3325–3335. (Citado na página 44.)

[17] Bugatti, P. H. (2008). Análise da influência de funções de distância para o processamento de consutas por similaridade em recuperação de imagens por conteúdo. Master's thesis, Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos. (Citado na página 49.)

[18] Bugatti, P. H., Traina, A. J. M., and Traina-Jr., C. (2008). Assessing the best integration between distance-function and image-feature to answer similarity queries. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1225–1230, Fortaleza, Ceara, Brazil. (Citado na página 49.)

[19] Castellano, G., Bonilha, L., Li, L., and Cendes, F. (2004). Texture analysis of medical images. *Clinical Radiology*, 59:1061–1069. (Citado na página 44.)

[20] Chong, S. Y. and Yao, X. (2007). Multiple choices and reputation in multiagent interactions. *IEEE Transactions on Evolutionary Computation*, 11(6):689–711. (Citado nas páginas 15 e 17.)

[21] Ciaccia, P., Patella, M., and Zezula, P. (1997). M-tree: An efficient access method for similarity search in metric spaces. In *International Conference on Very Large Databases (VLDB)*, pages 426–435, Athens, Greece. (Citado na página 48.)

[22] Congdon, P. (2006). *Bayesian Statistical Modelling*. Wiley Series in Probability and Statistics. (Citado na página 54.)

[23] Cordón, O., Herrera-Viedma, E., López-Puljalte, C., Luque, M., and Zarco, C. (2003). A review on the application of evolutionary computation to information retrieval. *International Journal of Approximate Reasoning*, 34:241–264. (Citado na página 49.)

[24] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40:5:1–5:59. (Citado nas páginas 2 e 42.)

[25] Daubechies, I. (1990). The wavelets transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36:961–1005. (Citado na página 44.)

[26] Deserno, T., Antani, S., and Long, R. (2009). Ontology of gaps in content-based image retrieval. *Journal of Digital Imaging*, 22:202–215. (Citado nas páginas 2 e 42.)

[27] Dimitrovski, I., Kocev, D., Loskovska, S., and Dzeroski, S. (2011). Hierarchical annotation of medical images. *Pattern Recognition*. DOI:10.1016/j.patcog.2011.03.026. (Citado na página 2.)

[28] Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31:198–211. (Citado na página 3.)

[29] Dorigo, M. and Caro, G. D. (1999). Ant colony optimization: A new meta-heuristic. *Proceedings of the Congress on Evolutionary Computation, IEEE Press*, 2:1470–1477. (Citado na página 26.)

[30] Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., second edition edition. (Citado nas páginas 10, 54 e 61.)

[31] Dy, J. G., Brodley, C. E., Kak, A., Broderick, L. S., and Aisen, A. M. (2003). Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):373–378. (Citado na página 50.)

[32] ElAlami, M. (2011). A novel image retrieval model based on the most relevant features. *Knowledge-Based Systems*, 24:23–32. (Citado na página 50.)

[33] Elder, J. F. and Pregibon, D. (1996). Advances in knowledge discovery and data mining. chapter A statistical perspective on knowledge discovery in databases, pages 83–113. American Association for Artificial Intelligence, Menlo Park, CA, USA. (Citado nas páginas 54 e 55.)

[34] Fernandez-Prieto, J., Canada-Bago, J., Gadeo-Martos, M., and Velasco, J. R. (2011). Optimisation of control parameters for genetic algorithms to test computer networks under realistic traffic loads. *Applied SoftComputing 2011*, to appear:9 pages. (Citado na página 39.)

[35] Fischer, B., Deserno, T. M., Ott, B., and Gunther, R. W. (2008). Integration of a research cbir system with ris and pacs for radiological routine. In *Proceedings of SPIE*, volume 6919, pages 691914–691914–10. (Citado nas páginas 2 e 42.)

[36] Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon. (Citado na página 26.)

[37] Freitas, A. A. (2001). Understanding the crucial role of attribute interaction in data mining. *Journal Artificial Intelligence Review*, 16(3):177–199. (Citado nas páginas 15 e 17.)

[38] Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Computer Graphics Image Processing*, 4:172–179. (Citado na página 44.)

[39] Gheyas, I. A. and Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition*, 43:5–13. (Citado na página 18.)

[40] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison Wesley. (Citado nas páginas 25, 26, 27, 28 e 37.)

[41] Graham, R., Perriss, R., and Scarsbrook, A. (2005). Dicom demystified: A review of digital file formats and their use in radiological practice. *Clinical Radiology*, 60:1133–1140. (Citado na página 2.)

[42] Guliato, D. and Rangayyan, R. (2011). Modeling and analysis of shape with applications in computer-aided diagnosis of breast cancer. *Synthesis Lectures on Biomedical Engineering*, 39(1):1–95. (Citado na página 3.)

[43] Guttman, A. (1984). R-tree : A dynamic index structure for spatial searching. In *ACM International Conference on Data Management (SIGMOD)*, pages 47–57, Boston, USA. ACM Press. (Citado na página 48.)

[44] Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the 17th Conference on Machine Learning*, pages 359–366, San Francisco, CA, USA. (Citado nas páginas 21 e 22.)

[45] Hamdani, T. M., Won, J., Alimi, A. M., and Karray, F. (2011). Hierarchical genetic algorithm with new evaluation function and bi-coded representation for the selection of features considering their confidence rate. *Applied Soft Computing*, 11:2501–2509. (Citado nas páginas 25 e 31.)

[46] Haralick, R. M., Shanmugam, K., and Deinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man. Cybern.*, 3(6):610–621. (Citado na página 44.)

[47] Haupt, R. L. and Haupt, S. E. (1998). *Practical Genetic Algorithms*. Wiley-Intercience. (Citado nas páginas 25, 26, 32, 36 e 37.)

[48] Haykin, S. (2009). *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition. (Citado nas páginas 54 e 60.)

[49] Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Michigan: MIT Press. (Citado nas páginas 27, 28 e 33.)

[50] Horng, J. and Yeh, C. (2000). Applying genetic algorithms to query optimization in document retrieval. *Information Processing & Management*, 36:737–759. (Citado na página 49.)

[51] Hruschka, E. R. and oes T. F., C. (2005). Feature selection for cluster analysis: an approach based on the simplified silhouette criterion. In *Proc. of the IEEE Int. Conf. on Computational Intelligence for Modelling and Automation*, pages 32–38, Vienna, Austria. (Citado nas páginas 7, 19 e 20.)

[52] Hsu, W., Lee, M., and Zhang, J. (2001). Image mining: trends and developments. *Journal of Intelligent Information Systems*, pages 7–23. (Citado nas páginas 42 e 61.)

[53] Huang, J., Kumar, S. R., Mitra, M., Zhu, W., and Zabih, R. (1997). Image indexing using color correlogram. In *IEEE International Conference on Computer Vision and Evolutionary Computation and Pattern Recognition*, pages 762–768, Puerto Rico. (Citado na página 43.)

[54] Ilyas, I. F., Beskales, G., and Soliman, M. A. (2008). A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40:11:1–11:58. (Citado na página 48.)

[55] Katayama, N. and Satoh, S. (2001). Distinctiveness-sensitive nearest neighbor search for efficient similarity retrieval of multimedia information. In *Proceedings of the 17th International Conference on Data Engineering (ICDE)*, pages 493–502, Washington, DC, USA. (Citado nas páginas 2, 11 e 12.)

[56] Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data - An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. (Citado na página 20.)

[57] Kim, W.-Y. and Kim, Y.-S. (2000). A region-based shape descriptor using zernike moments. *Signal Processing: Image Communication*, 16:95–102. (Citado na página 46.)

[58] Kinoshita, S. K., Azevedo-Marques, P. M. d., Pereira-Jr., R. R., Rodrigues, J. A. H., and Rangayyan, R. M. (2007). Content-based retrieval of mammograms using visual

features related to breast density patterns. *Journal of Digital Imaging*, 20(2):172–190. (Citado nas páginas 67 e 68.)

[59] Kiranyaz, S., Birinci, M., and Gabbouj, M. (2010). Perceptual color descriptor based on spatial distribution: A top-down approach. *Image and Vision Computing*, 28:1309–1326. (Citado na página 44.)

[60] Klami, A., Saunders, C., Campos, T. E., and Kaski, S. (2008). Can relevance of images be inferred from eye movements? In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 134–140. (Citado na página 49.)

[61] Korn, F., Pagel, B., and Faloutsos, C. (2001). On the 'dimensionality curse' and the 'self-similarity blessing'. *IEEE Trans. on Knowledge and Data Engineering*, 13(1):96–111. (Citado nas páginas 2 e 11.)

[62] Kudo, M. and Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41. (Citado nas páginas 5 e 19.)

[63] Li, L., Weinberg, C. R., A., D. T., and G., P. L. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Computerized Medical Imaging and Graphics*, 17(12):1131–1142. (Citado na página 19.)

[64] Liu, F. and Picard, R. W. (1996). Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. on Pattern Analysis and Machine Learning*, 18(7):184–189. (Citado na página 44.)

[65] Liu, H., Dougherty, E. R., Dy, J. G., Torkkola, K., Tuv, E., Peng, H., Ding, C., Long, F., Berens, M., Parsons, L., Zhao, Z., Yu, L., and Forman, G. (2005). Evolving feature selection. *IEEE Intelligent Systems*, 20:64–76. (Citado na página 19.)

[66] Liu, H., Hussain, F., Tan, C. L., and Dash, M. (2002). Discretization: An enabling

technique. *Data Mining and Knowledge Discovery*, 6(4):393–423. (Citado na página 22.)

[67] Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502. (Citado na página 19.)

[68] López-Pujalte, C., Guerrero-Bote, V., and Moya-Anegón, F. (2003a). Order-based fitness functions for genetic algorithms applied to relevance feedback. *Journal of the American Society for Information Science*, 54(2):152–160. (Citado na página 49.)

[69] López-Pujalte, C., Guerrero-Bote, V. P., and Moya-Anegón, F. (2003b). Order-based fitness functions for genetic algorithms applied to relevance feedback. *Journal of the American Society for Information Science*, 54(2):152–160. (Citado na página 53.)

[70] Lorena, A. C. (2006). *Investigação de estratégias para a geração de máquinas de vetores de suporte multiclasses*. Tese de doutorado, Instituto de Ciências Matemáticas e de Computação (ICMC/USP). (Citado na página 61.)

[71] Lu, J., Zhao, T., and Zhang, Y. (2008). Feature selection based-on genetic algorithm for image annotation. *Knowledge-Based Systems*, 21:887–891. (Citado nas páginas 19 e 50.)

[72] Longman Dictionary (2009). Longman dictionary of contemporary english. Pearson Education (DVD-ROM). Fifth Edition. (Citado na página 44.)

[73] Michalewicz, Z. (1992). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer Verlag. (Citado na página 31.)

[74] Michalewicz, Z. and Fogel, D. B. (2000). *How to solve it: modern heuristics*. Springer Verlag. (Citado nas páginas 39 e 40.)

[75] Mitchell, M. (1997). *An introduction to genetic algorithms*. Cambridge: MIT Press. (Citado na página 37.)

[76] Mitra, S. and Acharya, T. (2003). *Data Mining: Multimedia, Soft Computing and Bioinformatics*. John Wiley & Sons. (Citado na página 59.)

[77] Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *International Journal of Medical Informatics - IJMI*, 73(1):1–23. (Citado nas páginas 3 e 46.)

[78] Müller, H., Zhou, X., Depeursinge, A., Pitkanen, M., Iavindrasana, J., and Geissbuhler, A. (2007). Medical visual information retrieval: state of the art and challenges ahead. In *ICME International Conference, IEEE*, pages 683–686. (Citado na página 3.)

[79] Murray, S., Kersten, D., Olshausen, B., Schrater, P., and Woods, D. (2002). Shape perception reduces activity in human primary visual cortex. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 99, pages 15164–15169. (Citado na página 45.)

[80] Oliveira, G. M. B. (1999). *Dinâmica e Evolução de Autômatos Celulares Unidimensionais*. Tese de doutorado, Instituto Tecnologico de Aeronautica, Curso de Engenharia Eletronica e Computação na Área de Informática. (Citado nas páginas 30 e 38.)

[81] Oliveira, M., Cirne, W., and Azevedo-Marques, P. (2007). Towards applying content-based image retrieval in the clinical routine. *Future Generation Computer Systems*, 23:466–474. (Citado na página 3.)

[82] Op De Beeck, H., Torfs, K., and Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, 28:10111–10123. (Citado na página 45.)

[83] Paquerault, S., Hardy, P., Wersto, N., Chen, J., and Smith, R. (2010). Investigation of optimal use of computer-aided detection systems. the role of the “machine” in decision making process. *Academic Radiology*, 17(9):1112–1121. (Citado na página 3.)

[84] Pass, G., Zabih, R., and Miller, J. (1996). Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 65–73. (Citado na página 43.)

[85] Pedrycz, W., Park, B., and Pizzi, N. (2009). Identifying core sets of discriminatory features using particle swarm optimization. *Expert Systems with Applications*, 36:4610–4616. (Citado na página 18.)

[86] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238. (Citado na página 22.)

[87] Pourghassem, H. and Ghassemian, H. (2008). Content-based medical image classification using a new hierarchical merging scheme. *Computerized Medical Imaging and Graphics*, 32:651–661. (Citado na página 2.)

[88] Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco, USA. (Citado nas páginas 23, 54 e 57.)

[89] Rangayyan, R., Banik, S., and Desautels, J. (2010). Computer-aided detection of architectural distortion in prior mammograms of interval cancer. *Journal of Digital Imaging*, 23(5):611–631. (Citado na página 3.)

[90] Rangayyan, R. M., Desautels, J. E. L., and Ayre, F. J. (2011). Computer-aided diagnosis of breast cancer: Towards the detection of early and subtle signs. Teaching Files. Disponível em `http://enel.ucalgary.ca/People/Ranga/` (Acesso em 20/03/2011). (Citado na página 45.)

[91] Ribeiro, M. X., Balan, A. G. R., Felipe, J. C. Traina, A. J. M., and Traina-Jr (2009). Mining statistical association rules to select the most relevant medical image features. *Mining Complex Data, Springer Berlin / Heidelberg*, 165(1):113–131. (Citado na página 79.)

[92] Robnic-Sikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1-2):23–69. (Citado na página 22.)

[93] Samet, H. (2001). *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann. (Citado na página 49.)

[94] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press. (Citado na página 54.)

[95] Shannon, C. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27(1):379–423. (Citado na página 57.)

[96] Siedlecki, W. and Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10(5):335–347. (Citado na página 18.)

[97] Silva, S. F. (2007). Realimentação de relevância via algoritmos genéticos aplicada à recuperação de imagens. Master's thesis, Universidade Federal de Uberlândia. (Citado na página 31.)

[98] Silva, S. F., Barcelos, C. A. Z., and Batista., M. A. (2006a). The effects of fitness functions on genetic algorithms applied to relevance feedback in image retrieval. In *13th International Conference on Systems, Signals and Image Processing (IWSSIP'06)*, pages 443–446, Budapest, Hungary. (Citado na página 49.)

[99] Silva, S. F., Barcelos, C. A. Z., and Batista., M. A. (2006b). An image retrieval system adaptable to user's interests by the use of relevance feedback via genetic algorithm. In *XII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia'06)*, pages 45–52, Natal, RN. (Citado na página 49.)

[100] Silva, S. F., Barcelos, C. A. Z., and Batista, M. A. (2007a). Adaptive image retrieval through the use of a genetic algorithm. In *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 557–564, Patras, Greece. (Citado nas páginas 25 e 97.)

[101] Silva, S. F., Barcelos, C. A. Z., and Batista., M. A. (2007b). Adaptive image retrieval through the use of a genetic algorithm. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'2007)*, page 8p., Patras, Greece. (Citado na página 49.)

[102] Silva, S. F., Brandoli, B., Eler, D. M., Batista-Neto, J. E. S., and Traina, A. J. M. (2010). Silhouette-based feature selection for classification of medical images. In *Proceedings of the 23rd IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 315–320, Perth, Australia. (Citado nas páginas 92 e 97.)

[103] Silva, S. F., Ribeiro, M., Batista-Neto, J., Traina-Jr, C., and Traina, A. (2011). Improving the ranking quality of medical image retrieval using a genetic feature selection method. *Decision Support Systems*. (To appear):11 pages. doi:10.1016/j.dss.2011.01.015. (Citado nas páginas 2, 18, 49, 83 e 97.)

[104] Silva, S. F., Traina, A., Ribeiro, M., Batista-Neto, J., and Traina-Jr, C. (2009). Ranking evaluation functions to improve genetic feature selection in content-based image retrieval of mammograms. In *Proceedings of the 22rd IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–8, New Mexico, USA. (Citado nas páginas 25, 31, 49, 83 e 97.)

[105] Stehling, R. O., Nascimento, M. A., and Falcão, A. X. (2003). Cell histograms versus color histograms for image representation and retrieval. *Knowledge and Information Systems*, 5(3):315–336. (Citado na página 44.)

[106] Sun, J., Zhang, X., Cui, J., and Zhou, L. (2006). Image retrieval based on color distribution entropy. *Pattern Recognition Letters*, 27:1122–1126. (Citado na página 43.)

[107] Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32. (Citado na página 43.)

[108] Tahir, M. A., Bouridane, A., and Kurugollu, F. (2007). Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier. *Pattern Recognition Letters*, 28:438–446. (Citado na página 17.)

[109] Tamine, L., C., C., and Boughanem, M. (2003). Multiple query evaluation based on an enhanced genetic next term algorithm. *Information Processing & Management*, 39(2):215–231. (Citado na página 49.)

[110] Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-8(6):460–473. (Citado na página 44.)

[111] Tan, S. and Lewis, R. (2010). Picture archiving and communication systems: A multicentre survey of users experience and satisfaction. *European Journal of Radiology*, 75(3):406–410. (Citado na página 2.)

[112] Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press, New York, USA. (Citado nas páginas 10, 17 e 58.)

[113] Torres, R. S., Falcão, A. X., Gonçalves, M. A., Papa, J. P., B., Z., Fan, W., and Fox, E. A. (2009). A genetic programming framework for content-based image retrieval. *Pattern Recognition*, 42(2):283–292. (Citado na página 49.)

[114] Torres, R. S. and Falcão, A. X. (2006). Content-based image retrieval: Theory and applications. *RITA*, 13(2):165–189. (Citado na página 49.)

[115] Torres, R. S. and Falcão, A. X. (2007). Contour salience descriptors for effective image retrieval and analysis. *Image and Vision Computing*, 25(1):3–13. (Citado nas páginas 46 e 49.)

[116] Traina, A. J. M., Bueno, C. T. J. M., Chino, F. J. T., and Paulo, M. A. (2003). Efficient content-based image retrieval through metric histograms. *World Wide Web Journal*, 6(2):157–185. (Citado na página 44.)

[117] Traina, C., Traina, A. J. M., Faloutsos, C., and Seeger, B. (2002). Fast indexing and visualization of metric datasets using slim-trees. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 14(2):244–260. (Citado na página 48.)

[118] Vieira, S. M., Sousa, J., and Runkler, T. (2010). Two cooperative ant colonies for feature selection using fuzzy models. *Expert Systems with Applications*, 37:2714–2723. (Citado na página 18.)

[119] Volnyansky, I. and Pestov, V. (2009). Curse of dimensionality in pivot based indexes. In *Proceedings of the 2009 Second International Workshop on Similarity Search and Applications*, pages 39–46. (Citado na página 11.)

[120] Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, United States, second edition edition. (Citado na página 22.)

[121] Yadav, R. B., Nishchal, N. K., Gupta, A. K., and Rastogi, V. K. (2007). Retrieval and classification of shape-based objects using fourier, generic fourier, and wavelet-fourier descriptors technique: A comparative study. *Optics and Lasers in Engineering*, 45:695–708. (Citado na página 46.)

[122] Yan, H., Zheng, J., Jiang, Y., Peng, C., and Xiao, S. (2008). Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. *Applied Soft Computing*, 8:1105–1111. (Citado na página 18.)

[123] Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224. (Citado na página 22.)

[124] Zezula, P., Amato, G., Dohnal, V., and Batko, M. (2006). *Similarity Search: The Metric Space Approach*, volume 32. Springer: Series Advances in Database Systems. (Citado na página 49.)

[125] Zhang, D. and Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37:1–19. (Citado na página 46.)

[126] Zhao, T., Lu, J., Zhang, Y., and Xiao, Q. (2008). Feature selection based on genetic algorithm for cbir. In *IEEE Congress on Image and Signal Processing*, volume 2, pages 495–499. (Citado na página 19.)

[127] Zhao, Z. and Liu, H. (2009). Searching for interacting features in subset selection. *Intelligent Data Analysis archive*, 13(2):207–228. (Citado nas páginas 15 e 17.)

[128] Zhu, Z., Chen, X., Zhu, Q., and Xie, Q. (2007a). A ga-based query optimization method for web information retrieval. *Applied Mathematics and Computation*, 185:919–930. (Citado na página 49.)

[129] Zhu, Z., Ong, Y., and Dash, M. (2007b). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40:3236–3248. (Citado na página 23.)

[130] Zhu, Z., Ong, Y., and Dash, M. (2007c). Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Trans. on Systems Man, and Cybernetic*, 37(1):70–6. (Citado na página 23.)