CONEM 2012 «Engenharia em destaque»
VII Congresso Nacional de Engenharia Mecânica
São Luís - Maranhão - Brasil
31 de julho a 03 de agosto
www.abcm.org.br/conem2012

TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS A IMAGENS TÉRMICAS MASTOLÓGICAS

Giomar Oliver Sequeiros Olivera, goliver@ic.uff.br[1]
Tiago Bonini Borchartt, tbonini@ic.uff.br [1]
Aura Conci, aconci@ic.uff.br[1]
Leandro Augusto Frata Fernandes, laffernandes@ic.uff.br [1]
Rita de Cássia Fernandes de Lima, ritalima@ufpe.br[2]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF), Rua Passo da Pátria 156 - Bloco E - 3º andar , São Domingos Niterói - RJ - CEP: 24210-240 – Brazil.
[2]Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235 - Cidade Universitária, Recife – PE

Resumo: *Imagens térmicas vêm sendo usadas há várias décadas com o propósito de supervisionar a distribuição da temperatura no corpo. Disto segue uma enorme base de imagens que precisam ser consultadas e recuperadas. Este trabalho apresenta uma utilização modificada do processo de descoberta de conhecimento em banco de dados (knowledge discovery in databases ou KDD), orientado ao tratamento de imagens térmica e aplicado sobre o banco de imagens do projeto CAPES-PROENG 021/2008 desenvolvido na UFF. Esse projeto objetiva usar imagens térmicas da mama na detecção de patologias. Primeiramente selecionam-se imagens com diagnóstico conhecido. A seguir essas imagens são convertidas de pseudocores a tons de cinza, sendo depois segmentadas em duas regiões de interesse. O passo seguinte é o processo de extração de características baseadas em medidas estatísticas (média, desvio padrão, curtose, entre outras), dispostas em um vetor de características associado às imagens. Depois é feita a integração e transformação das características, criando-se assim uma nova base de dados, onde são aplicadas técnicas de mineração, como a classificação baseada em árvores de decisão, máquina de vetores de suporte, Naive Bayes e redes neurais. A avaliação dos resultados é feita mediante o uso das medidas: acurácia, sensibilidade, especificidade e área sob as curvas ROC (receiver operating characteristic). Também são usadas técnicas de clusterização (e.g., k-means, x-means e dbscan), que fazem uso da análise do centróide e número de clusters encontrados. Finalmente, quando necessário, é aplicado um processo de refinamento para aprimorar os resultados. O conhecimento gerado pode ser de grande utilidade para dignóstico médico, auxiliando o futuro desenvolvimento de um sistema de Computer Aided Diagnosis (CAD) baseado nas características extraídas.*

Palavras-chave: mineração de imagens, imagens térmicas, classificação, auxílio ao diagnóstico, processo kdd

1. INTRODUÇÃO

O câncer de mama é o segundo câncer mais freqüente no mundo e o que apresenta o maior número de casos entre mulheres no Brasil (INCA, 2012). Atualmente a mamografia é o método padrão para a detecção de tumores por possuir uma alta precisão. Este tipo de exame, entretanto, apresenta algumas desvantagens como expor a paciente a uma dose de radiação, ter dificuldades na detecção de tumores no caso de mamas mais densas, além de apresentar baixa capacidade de detecção de patologias nos primeiros estágios de desenvolvimento. Estas limitações da mamografia, juntamente com o crescente número de casos de câncer em pacientes mais jovens, motiva o desenvolvimento de novas técnicas para a detecção precoce de patologias na mama, dentre as quais, destaca-se a termografia.

Termografia é um exame fisiológico, não invasivo e sem uso de radiações ionizantes. Possibilita a detecção de tumores mamários, muito antes que qualquer outro método, ainda quando as células produzem substâncias responsáveis pela criação de neovascularização que "alimentará" o futuro tumor. Em conseqüência, constata-se que na região onde há uma patologia mamária a temperatura da pele pode ser maior do que a temperatura do tecido normal. Com isso, pode ocorrer uma assimetria na distribuição das temperaturas das mamas, então é possível que uma delas possa apresentar algum tipo de patologia (Serrano, 2010).

Imagens térmicas ou termogramas são adquiridos por uma câmera sensível à radiação infravermelha (i.e., calor). O baixo custo e a facilidade de obtenção deste tipo de imagens criam a necessidade de armazenamento em bancos de dados para sua posterior análise e gerenciamento. Para extrair conhecimento dessa base são necessários sistemas e algoritmos que possam ser aplicados a um grande volume de dados e ao mesmo tempo possam encontrar, de maneira eficiente, padrões e relacionamentos que estejam ocultos. Assim, torna-se clara a utilidade da mineração de dados em bancos de imagens, pois esta propõe técnicas para extrair conhecimento não-explícito (i.e., relacionamentos espaciais, padrões relevantes, dentre outros) a partir de grandes acervos de imagens. Este artigo aborda temas fundamentais desta área de pesquisa.

Organiza-se este trabalho como segue: na Seção 2 serão apresentados os conceitos da mineração de dados e trabalhos relacionados com a mineração de imagens. A Seção 3 apresenta a metodologia proposta e aplicada, desde a seleção de imagens, depois o pré-processamento, até a aplicação de algoritmos e técnicas de extração de conhecimento. Na seqüência, na Seção 4, serão apresentados e avaliados os resultados obtidos. Finalmente, na Seção 5, serão apresentadas as conclusões e direções para trabalhos futuros.

2. MINERAÇÃO DE IMAGENS

A descoberta de conhecimento em banco de dados (*knowledge discovery in databases* ou KDD) consiste na aplicação de várias etapas de processamento instruído com algumas decisões tomadas pelo usuário. Uma visão prática da aplicação de KDD, enfatizando a interatividade do processo, é mostrada na Figura 1 (Fayyad et a.l, 1996).

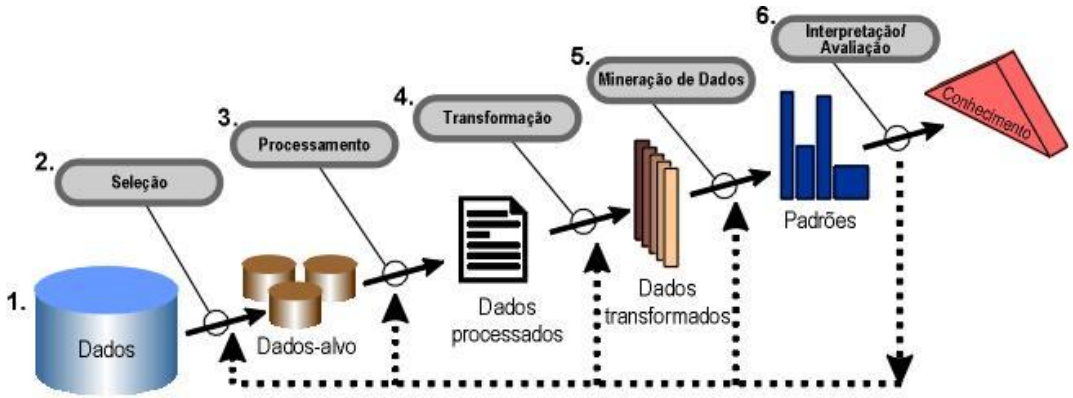

Figura 1: Etapas do processo KDD (Fayyad et al., 1996)

As etapas numeradas na Figura 1 podem ser detalhadas por:

1. Banco de dados - representa o conjunto contendo a informação a ser minerada.
2. Seleção - de acordo com os objetivos desejados, um subconjunto dos dados disponíveis é selecionado para a etapa de descoberta de conhecimento.
3. Processamento - os dados são processados e passam por uma limpeza onde há a integração de dados heterogêneos, eliminação de dados incompletos, as repetições e problemas na definição de tipos.
4. Os dados são transformados para serem armazenados adequadamente, visando facilitar a aplicação dos algoritmos de mineração.
5. Mineração de dados é a fase principal do processo KDD, onde se escolhe a tarefa de mineração a ser aplicada (classificação, clusterização, regras de associação, sumarização, etc.), dependendo do objetivo que se deseja alcançar.
6. Na etapa de interpretação/avaliação os resultados obtidos são interpretados e ranqueados para serem apresentados para o usuário.

A Mineração de Imagens (MI) é uma área multidisciplinar (Figura 2) relacionado com várias outras áreas do conhecimento, tais como: Visão Computacional, Processamento de Imagens, Mineração de Dados, Bancos de Dados, Aprendizado de Máquina e Sistemas de Recuperação de Imagens baseados no conteúdo das mesmas (Vieira et al., 2002). Assim como na Mineração de Dados existe uma fase de seleção e obtenção de atributos a partir de várias tabelas ou bases de dados (Fayyad et al., 1996) para gerar as instâncias a serem avaliadas. Na MI a obtenção dessas características também utiliza técnicas de Processamento de Imagens (Gonzáles; Woods, 2000), tratando da extração de conhecimento implícito, relacionamento entre dados de imagens, ou outros padrões não explicitamente armazenados em banco de dados de imagens. Seu foco é determinar como a representação de baixo nível (e.g., pixels) de uma imagem pode ser eficiente e efetivamente processada para identificar objetos e seus relacionamentos (Zhang et al., 2001).

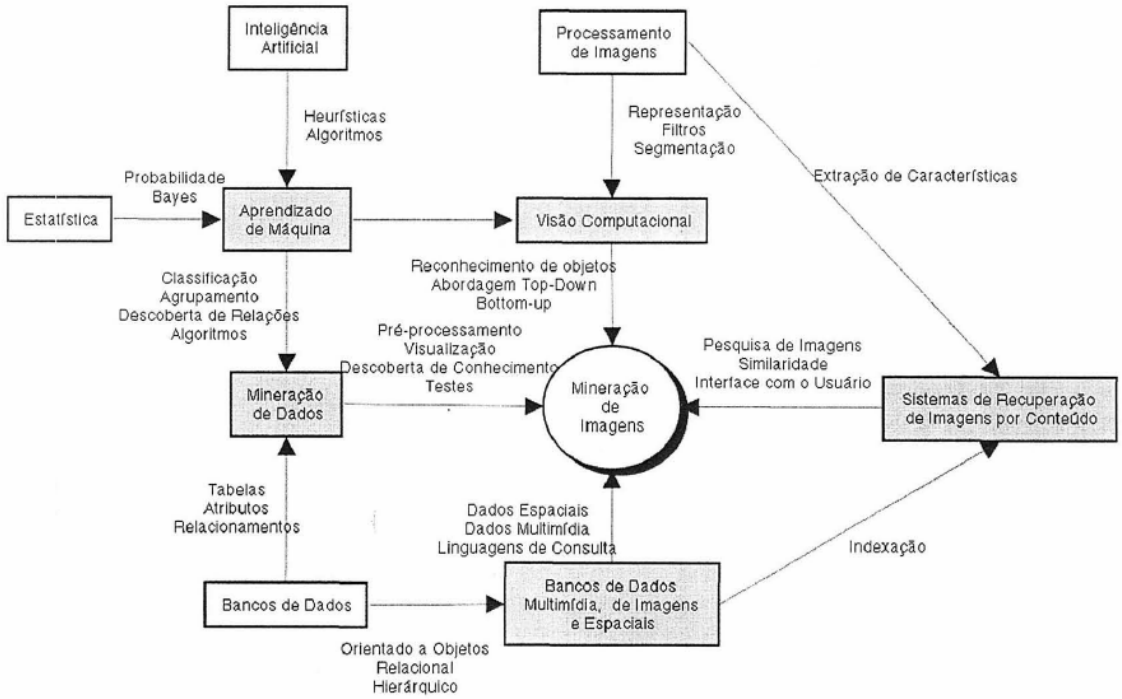

Figura 2: Integração da Mineração de Imagens com outras disciplinas (Vieira et al. , 2002)

2.1. Classificação

Classificar consiste em associar um conjunto de atributos (tupla) a um atributo objetivo chamada de classe. De maneira mais formal, a tarefa de classificação trata de "descobrir algum tipo de relacionamento entre valores de predição e o valor objetivo, tal que o conhecimento descoberto possa ser usado para prever a classe de outra tupla" (Maimon; Rokach, 2010).

Árvore de Decisão é uma técnica bastante utilizada em tarefas de classificação. Para estimar a idade gestacional (IG) através de imagens digitais da superfície plantar de recém-nascidos, Araújo et al. (2007) utilizaram algoritmos de J.48, M5P (Quinlan, 1993), REPTree e o LMT (Landwehr et al., 2003), implementados no software WEKA (Hall et al., 2009). A fase de pré-processamento foi um passo importante para gerar um banco de dados que auxilia em como fazer a mineração. O resultado final foi um sistema que mantém um banco de dados de conhecimento que evolui conforme novas instâncias e características são adicionadas a ele. Quando uma nova imagem é analisada, suas características são extraídas e os modelos de mineração mencionados são aplicados para a instância corrente, gerando dois tipos de saída: uma distribuição de probabilidades por IG; e um escore exato para a IG. A precisão dos classificadores foi similar. Posteriormente o mesmo sistema é ampliado usando árvores de decisão Fuzzy, que fornecem uma distribuição de probabilidades mais concisa sobre as IG possíveis para cada neonato. Esses algoritmos são aplicados a base de características extraídas de imagens, fornecendo várias informações para auxiliar no cálculo final da IG que, por fim, é validado por um especialista (Araújo, 2006)**.** Outra abordagem baseada em árvores de decisão é o SKYCAT (Fayyad et al., 1996 - a), originário de um sistema que provê um ambiente integrado para a construção, classificação, gerência e análise de catálogos de imagens astronômicas e que utiliza variações dos algoritmos ID3 e C4.5, GID3 e O-BTree. O algoritmo, GID3, foi utilizado para diminuir o particionamento dos dados, criando subdivisões que representam mais de um valor. O outro algoritmo, O-Btree, utiliza outras medidas para seleção dos dados, ao invés de usar apenas o grau de entropia.

A técnica de classificação de Naive Bayes baseada em probabilidades condicionais (Witten; Frank, 2000) é usada para minerar imagens pelo sistema GeoBrowse (Kopersi; Marchisio, 2000), que provê uma interface para consultas por similaridade e um módulo de análise de clusterização de imagens de satélites. Para cada imagem, são extraídos três tipos de vetores de características, conforme a classificação do dado: nível de pixel (cor, textura); nível de região (contorno, forma, tamanho, dentre outros atributos) e nível de faixa (*tile*, semelhante à região, porém em relação à imagem inteira). Um exemplo de mineração de imagens envolvendo redes Bayesianas (Vailaya et al., 2006), onde há a classificação hierárquica de imagens fotográficas diversas, é apresentado objetivando treinar o classificador para avaliar as imagens segundo conceitos de alto nível, como fotografias de cenas internas/externas, cidade/paisagem e pôr-do-sol/floresta/montanha.

Existem ainda técnicas baseadas em redes neurais artificiais (*Artificial Neural Networks*). Antonie et al. (2001) aplicam um classificador baseado em redes de neurais e apresentam alguns experimentos para a detecção de tumores em mamografias digitais a partir do uso do algoritmo de *back-propagation* para treinar as redes, e classificar as

mamografias em duas categorias: normais e anormais. As normais são aquelas que caracterizam um paciente saudável. As anormais incluem também os casos de tumores benignos e os casos malignos, i.e. mamografias tomadas de pacientes com tumores. O pré-processamento das imagens foi fundamental para obter uma boa acurácia nos resultados finais atingindo o 70% dos casos. Uma característica das redes neurais é o elevado tempo de processamento que requer, porém os resultados são bons. O uso de redes neurais mais sofisticadas pode-se ter um ganho no tempo e na acurácia. Outra abordagem que também trabalha com imagens mamográficas é apresentada por Ferrero (2005). Arora et al. (2008) usam termogramas das mamas de 94 pacientes, sendo 60 com tumor maligno e 34 com tumor benigno, onde são extraídas aproximadamente 100 imagens por cada paciente em uma classificação (diagnóstico de tumor benigno e tumor maligno) usando uma rede neural artificial. Os resultados reportados são de 61,70% de acurácia, 96,70% de sensibilidade e 26,50% de especificidade.

A classificação usando máquina de vetores de suporte (*support vector machine* ou SVM) mostra bons resultados nos banco de dados de imagens por esses serem compostos por atributos numéricos. Em Mert (2001) um banco de dados de imagens médicas obtidas mediante o exame de PAAF (punção aspirativa com agulha fina) é usado para classificar tumores como benignos ou malignos. Este classificador para o auxílio ao diagnóstico por imagem térmica de patologias de mama também é usado em Resmini (2011) na rotulação de imagens como "com patologia" ou "sem patologia" para isso são usadas as imagens segmentadas da mama esquerda e direita de 28 pacientes (24 com patologia e 4 sem patologia), sendo reportada uma acurácia de 82,14%, sensibilidade de 91,70%, especificidade de 25%, e área abaixo da curva ROC (*receiver operating characteristic*) de 0,58. As características extraídas das imagens foram: medidas estatísticas, fractais e geoestatísticas. Uma abordagem similar com imagens térmicas é tratada em Borchartt et al. (2011)**.** Outras abordagens para a classificação de imagens médicas são vistos em Eddaoudi (2011), Rolando (2006) e Dheeba (2010). Um ponto a levar em conta é que ao se tratar de imagens médicas o mais importante é a precisão dos resultados porque as decisões dos especialistas são críticas.

2.2. Descoberta de Regras de Associação

Uma regra de associação é uma expressão $X \rightarrow Y$ aplicada sobre um conjunto de transações D, implicando que a cada transação T que contenha X, seja provável conter Y, respeitando critérios de confiabilidade e suporte (Hipp et al., 2000). Uma aplicação que envolve descoberta de regras de associação para um sistema de recuperação de imagens com um módulo para diagnosticar câncer de pele pode ser vista em Wang e Chung (2001). Em Ordonez e Omiecinski (1999) é apresentado um sistema para busca de regras de associação em imagens coloridas, que segue os passos de extração de características, identificação de objetos, criação de registros, geração de imagens auxiliares e finalmente a mineração de imagens.

Em Djeraba (2001) é proposto um sistema semelhante para descoberta de relacionamentos escondidos através de características de imagens agrupadas em dicionários. Combinando soluções de recuperação de imagens por conteúdo e aprendizado de máquina, a extração de características combina o que o autor chama de *agrupamento simbólico* e a descoberta de relacionamentos propriamente dita. A primeira etapa se baseia em uma fase de pré-processamento, na qual é feita a montagem de um dicionário de características, que é o agrupamento das características mais semelhantes. Na seqüência, para cada grupo de características de uma imagem, é associada uma representação simbólica baseada no dicionário para simplificar a fase de descoberta de relacionamentos.

O algoritmo MaxOcurr, uma modificação do algoritmo Apriori, é usado por Zaïane e Han (2000) para extrair regras com itens repetidos a partir de uma base de dados espaciais. Uma das observações é que dados visuais possuem propriedades específicas do domínio de imagens, de modo que algumas características visuais podem estar repetidas em uma mesma imagem e essas ocorrências podem ser consideradas relevantes. O processo pode ser resumido em três passos. Primeiro é feita a segmentação de regiões baseada em cor e textura, dando origens a áreas que compartilham algumas características, tais como massa, variância e centróide. Depois a identificação de relações espaciais nas áreas extraídas, tais como *disjoint, inside, contains, equals, meets, covered by e overlaps*, usando diferentes resoluções da imagem, para melhoria de desempenho. Finalmente é feita a aplicação do algoritmo de extração de relações.

É comum que a mineração de regras de associação seja como um método de apoio ao diagnóstico médico (CAD), como sugestão automática ou de laudo (segunda opinião). Para isso é preciso relacionar os dados de baixo nível (as características da imagem) com os de alto nível (os jargões do especialista). A Figura 3 exemplifica esses detalhes. Outras aplicações se têm em Yi et al. (2005)**,** onde é mostrado que as regras de associação podem estabelecer automaticamente a associação semântica entre as imagens de consulta e as imagens de *feedback* que um usuário faz num sistema de recuperação de imagens por conteúdo. Em Ishwar (2001) se tem outra abordagem para sistemas de recuperação de imagens. As regras de associação são usadas para ajudar a detecção de câncer em mamografias (Zaïane, 2002). O método NFARM (Novel Fuzzy Association Rule Mining) para detectar tumores no cérebro é apresentado em Rajendran (2010).

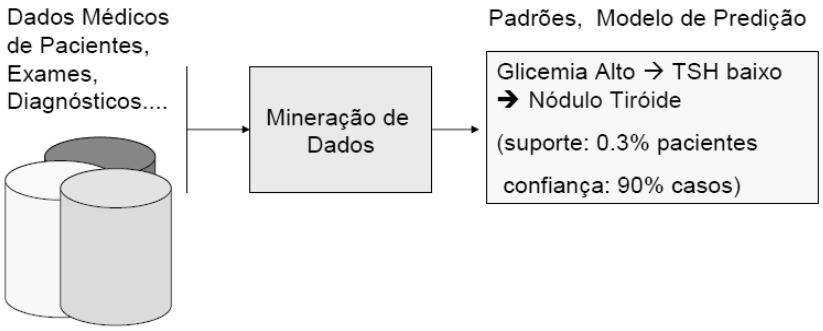

Figura 3. Dados de baixo nível e alto nível (Yi et al., 2005)

2.3. Clusterização

Na atividade de clusterização o sistema de KDD deve induzir, de maneira não supervisionada, um esquema de agrupamento que particiona o conjunto de treinamento em classes (Celinski; Bellon, 1998). Para análise de agrupamentos podem ser usados algoritmos específicos, que podem ser classificados em: hierárquicos, aglomerativos, divisivos, não-hierárquicos e de partição (Carvalho, 2001). Esses algoritmos têm como objetivo classificar, com respeito a algum critério pré-determinado, uma amostra de entidades em grupos mutuamente exclusivos baseado nas suas similaridades. O que define algoritmos hierárquicos ou divisivos é que a união de dois grupos numa certa etapa produz um dos grupos da etapa superior, formando uma hierarquia.

O algoritmo COBWEB/3 proposto em Soh e Tsatsoulis (1999) agrupa regiões por características espectrais, espaciais e de textura de imagens de satélites. Nesse caso, a mineração de imagens é usada como um instrumento auxiliar ao processo de visão computacional. O funcionamento do COBWEB/3 consiste na formação de conceitos probabilísticos, no qual uma instância pertence a um conceito com certo grau de certeza. Em resumo, a cada iteração, um algoritmo desse tipo agrupa as instâncias em conceitos, sumariza-as e organiza-as hierarquicamente.

O algoritmo k-means também é aplicado em Samma (2009), Siddheswar e Rose (1999) e Narasimhan (2009) na segmentação de imagens ou como uma medida de avaliação simples baseado no intra e inter-cluster que permite que o número de clusters seja determinado automaticamente no algoritmo k-means. A principal desvantagem do k-means é que o número de clusters, k, deve ser fornecido como um parâmetro e, tratando-se de imagens, é difícil saber o número de clusters que se deseja gerar. Outros exemplos que envolvem agrupamentos estão disponíveis em Zhang et al. (2001). A maioria destes está envolvida em estágios iniciais do processo de mineração. Em geral, após a execução do algoritmo, é necessária a intervenção humana para rotulação dos grupos gerados.

3. METODOLOGIA

A metodologia proposta, similarmente ao processo KDD descrito na Seção 2, é composta de seis (6) passos. No primeiro passo é feita a obtenção das imagens térmicas seguindo um protocolo padrão. No segundo passo é feito um pré-processamento, onde se convertem as imagens a escalas de cinza e é feita a segmentação da região de interesse. No terceiro passo é feita a extração das características das imagens. O quarto considera a transformação dos dados gerando um arquivo compatível com o software de mineração. No quinto passo são usados os algoritmos de mineração de dados como classificação, extração de regras de associação e clusterização nas características extraídas no passo anterior. Finalmente no sexto passo é feita a avaliação e interpretação dos resultados.

3.1. Imagens utilizadas

Projeto Pró-Engenharias 021/2008 (Proeng), aprovado em 2008, envolve o Instituto de Computação da Universidade Federal Fluminense (UFF) e o Departamento de Engenharia Mecânica da Universidade Federal de Pernambuco (UFPE) na aquisição de imagens térmicas como a da Figura 4. Essas imagens são adquiridas no Hospital das Clínicas da Universidade Federal de Pernambuco (HC-UFPE) usando uma câmera Flir Thermacam, modelo S45. Os pacientes que são submetidos ao exame assinam um termo de consentimento livre e esclarecido liberando as imagens para estudo. Foram usadas 28 imagens, das quais 24 são de pacientes com alguma doença na mama e 4 de pacientes sem qualquer doença mamária. Posteriormente foram acrescentadas 6 novas imagens de pacientes sem patologia para melhor balanceamento da base de dados.

3.2. Pré-processamento e segmentação

As imagens utilizadas têm dimensões de 320x240 pixels. O software FLIR Quick View mapeia as temperaturas medidas para uma faixa de tons de cores, sendo possível usar diversas paletas disponíveis. No pré-processamento a paleta de representação das temperaturas é mudada para a paleta em tons de cinza. A segmentação inicial usa o método

automático de Motta (2010) resumido na Figura 5. Esse consiste em pegar a imagem termográfica em tons de cinza e determinar o ponto de corte inferior da imagem usando a região com a maior quantidade de pixels brancos abaixo da mama por algoritmos de limiarização e morfologia matemática (e.g., dilatação e erosão). A seguir é definido o ponto de corte superior usando as axilas como limites superiores. A axila é encontrada usando a intersecção de uma linha horizontal e os limites entre o corpo e o fundo. Depois se faz a remoção das regiões que não pertencem à região de interesse (ROI), como os braços, usando o algoritmo *Flood Fill*. Em seguida, usa-se a transformada de Hough para a detecção das pregas inframamárias, caracterizadas como a maior parábola da imagem que corresponde ao limite inferior da mama. Finalmente é feita a separação das mamas em áreas iguais e um deslocamento vertical, se necessário, para que os limites inferiores das duas ROIs correspondam ao limite inferior da mama. Ainda é feito uma nova segmentação manual das ROI das mamas esquerda e direita para retirar áreas que correspondem ao corpo das pacientes, mas não de regiões das mamas. A Figura 6 mostra o resultado do processo.

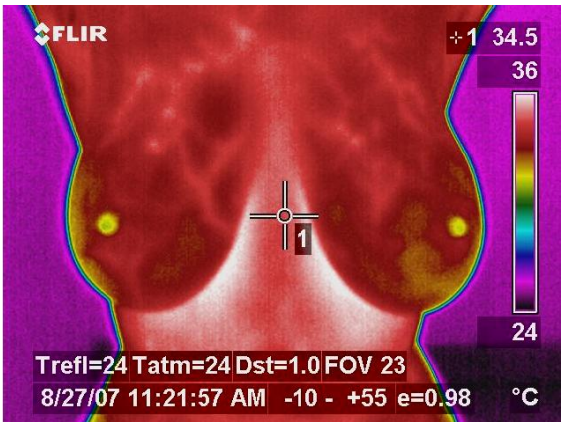

Figura 4: Termografia de uma paciente (Proeng).

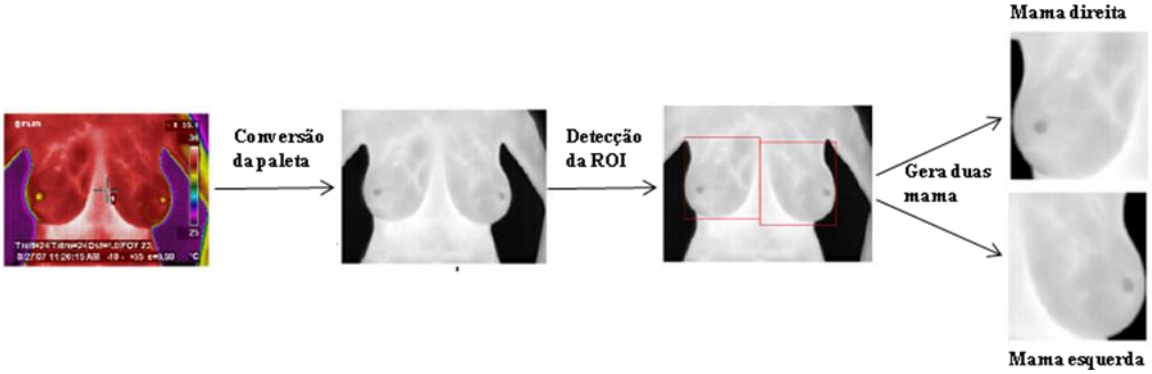

Figura 5: Algoritmo de segmentação automática (Motta, 2010).

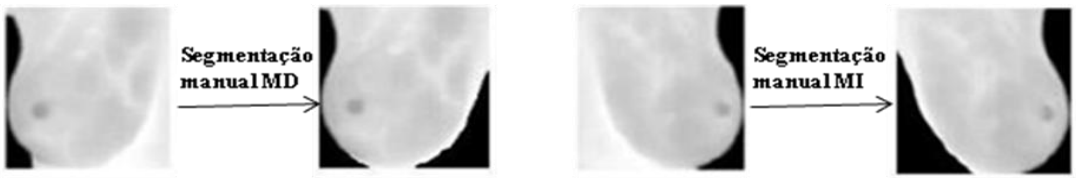

Figura 6: Segmentação manual

3.3. Extração de características

O software ImageJ foi usado para a extração das características: média, desvio padrão, tom mínimo e máximo, mediana, assimetria e curtose da imagem baseados no histograma em escala de cinza da ROI. As definições formais destas características são facilmente disponíveis na literatura (Gonzáles; Woods, 2000). Estas medidas estatísticas geram 7 características para cada mama, resultando um total de 14 características por imagem de uma paciente.

3.4. Transformação

O software WEKA - *Waikato Environment for Knowledge Analysis* (Hall et al., 2009)**,** desenvolvido pela Universidade de Waikato - Nova Zelândia, tem por objetivo reunir implementações de algoritmos de mineração, dentre eles: (1) o algoritmo de classificação por máquina de vetores de suporte chamado de SMO no WEKA; (2) o algoritmo C4.5, chamado de J48; (3) o algoritmo de classificação usando redes de neurais artificiais chamado de Multilayer Perceptron; (4) algoritmos baseados em Naive Bayes; (5) algoritmos de clusterização como o k-means; e (6) o APriori,

para a descoberta de regras de associação. Para usar esta ferramenta foi necessário fazer uma conversão ao formato ARFF das características criando um arquivo de texto para seu processamento. Além disso, foi necessário adicionar um atributo chamado de "Classe" que toma os valores de "C" e "S" (com patologia e sem patologia respectivamente) que serve para o treinamento da base de dados e a avaliação dos classificadores.

3.5. Mineração de dados

Utilizou-se uma técnica de validação cruzada estratificada de 10 (parâmetro Folds=10). A validação cruzada consiste em, dado um número k, dividir os dados em k subconjuntos mutuamente exclusivos do mesmo tamanho. A partir disto, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para o treinamento do classificador. Este processo é repetido durante k iterações com cada um dos subconjuntos de teste. Finalmente, calcula-se a média aritmética dos resultados de cada iteração para obter um único resultado.

Foram utilizados quatro algoritmos de classificação com os parâmetros default do software WEKA, como segue:

1. Máquina de vetores de suporte: weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 –K "weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"
2. Árvore de decisão J48: weka.classifiers.trees.J48 -C 0.25 -M 2
3. Redes neurais: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
4. Naive Bayes: weka.classifiers.bayes.NaiveBayes

Para fazer a clusterização foram usados os algoritmos k-means, X-means e DBScan. Ao todo foram utilizadas 34 imagens com a seguinte parametrização da ferramenta:

1. K-means: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10 Con k=2 (clusters ou grupos)
2. X-means: weka.clusterers.XMeans -I 1 -M 1000 -J 1000 -L 2 -H 4 -B 1.0 -C 0.5-D Weka.core.EuclideanDistance -R first-last" -S 10
3. DBScan: weka.clusterers.XMeans -I 1 -M 1000 -J 1000 -L 2 -H 4 -B 1.0 -C 0.5-D Weka.core.EuclideanDistance -R first-last" -S 10

3.6. Interpretação e avaliação

Para a avaliação dos classificadores foram usadas as medidas padrão de acurácia, sensibilidade, especificidade e curvas ROC apresentadas nas Equações (1), (2) e (3), onde VP, VN, FP e FN são respectivamente, os números de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

$$Acurácia = \frac{VP + FP}{VP + FP + VN + FN} \tag{1}$$

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{3}$$

Curvas ROC definem um método gráfico bidimensional para avaliação, organização e seleção de sistemas de diagnostico ou predição. Nelas, o eixo horizontal representa os valores da taxa de falsos positivos (1− especificidade) e o eixo vertical os valores de sensibilidade. Para quantificar uma análise usando a curva ROC pode-se calcular a área abaixo da curva, sendo que quanto mais próximo da unidade for o valor, melhor será a metodologia usada.

4. RESULTADOS

Na Tabela 1 são apresentados os resultados médios obtidos pelos os algoritmos de classificação do software WEKA. Pode-se observar que embora as curvas ROC dos métodos empregados sejam baixas, os classificadores obtiveram uma acurácia razoável, especialmente o Naive Bayes, onde essa atingiu um valor de 82,7%.

Tabela 1. Resultados obtidos dos classificadores com 28 imagens.

Método	Sensibilidade (%)	Especificidade (%)	Acurácia (%)	Curva ROC (área)
Máquina de vetores de suporte	85,7	79,1	73,5	0,500
J48	75,0	73,5	72,0	0,349
Redes Neurais	75,0	76,1	77,4	0,625
Naive Bayes	78,6	80,2	82,7	0,580

Na Tabela 2, os resultados de outra abordagem (Resmini, 2011) são mostrados para comparação. Nesta análise as mesmas imagens são utilizadas, mas empregam-se mais características: 4 medidas estatísticas simples (range, média, desvio padrão e quantização do último tom da posterização de 8 tons), dimensão fractal de Higuchi e medidas

geoestatísticas (coeficiente de Geary e índice de Morán). Além disso, a ROI é dividida em 4 partes, de modo que tem-se um total de 712 características por paciente. Mas, no trabalho usado para comparação, apenas o SVM é usado como classificador.

Tabela 2. Resultados da comparação com Resmini (2011).

Método	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
Todas as medidas	91,70	25,00	82,14
Medidas estatísticas	83,33	25,00	75,00
Fractais	79,16	0,00	67,85
Medidas geoestatísticas	91,67	25,00	82,14

Comparando-se a primeira linha da Tabela 1 com a Tabela 2, observa-se que as medidas estatísticas não apresentam bons resultados em comparação com outras características, em especial com as medidas geoestatísticas. Observa-se também que SVM é uma técnica com ótima sensibilidade. Além existe a necessidade de se ter uma base de dados mais balanceada quanto ao número de pacientes com patologia e sem patologia.

Uma característica do processo KDD é que quando não se atingem resultados bons, pode-se voltar a qualquer etapa anterior e fazer alguma mudança. Neste caso, na etapa de seleção foi necessário acrescentar mais 6 imagens de pacientes sem patologia para equilibrar o banco de dados somando 34 imagens em total. Na Tabela 3 pode-se observar que todos os métodos obtiveram um ganho nos valores de sensibilidade, especificidade, acurácia e curvas ROC, demonstrando assim a importância da etapa de seleção de dados. O classificador Naive Bayes obteve um melhor resultado na área da curva ROC que ficou mais próximo à unidade, seguido do classificador de Redes Neurais.

Tabela 3. Resultados obtidos dos classificadores com 34 imagens.

Método	Sensibilidade (%)	Especificidade (%)	Acurácia (%)	Curva ROC (área)
Máquina de vetores de suporte	85,3	84,4	85,3	0,779
J48	79,4	78,2	78,5	0,800
Redes Neurais	76,5	77,0	78,1	0,846
Naive Bayes	79,4	80,1	82,0	0,892

A Tabela 4 mostra os resultados depois de aplicar o algoritmo k-means, X-means e DBScan na clusterização não supervisionada (sem conhecer a qual classe pertence, com patologia ou sem patologia). Observe-se que a distribuição dos dados é diferente para cada técnica. No k-means o cluster 1 tem 18 instâncias e o cluster 2 tem 16 (o qual é uma distribuição não tão boa tendo em conta que a distribuição das classes são de 24 e 10 instâncias). Um melhor resultado foi obtido pelo X-means que teve uma aproximação de 25 e 9 instâncias, respectivamente. Por não precisar de um número de clusters como parâmetro de entrada, o algoritmo DBScan não conseguiu clusterizar as instâncias dando como resultado um cluster único com 2 instâncias.

Tabela 4. Resultados obtidos da clusterização.

Método	Cluster 1	Cluster 2
K-Means	18 (53%)	16 (47%)
X-Means	9 (26%)	25 (74%)
DBScan	32 (100%)	2 sem clusterizar

5. CONCLUSÕES E TRABALHOS FUTUROS

Neste artigo é descrito como as técnicas de mineração de dados podem ser aplicadas à mineração de imagens térmicas com a finalidade de extrair informações implícitas para auxílio na tomada de decisões. Foram usadas 34 imagens térmicas com diagnóstico conhecidos e empregados como classificadores na máquina de vetores de suporte (SVM), árvores de decisão (J48), *Naive Bayes* e redes neurais. Destes, o SVM apresentou melhores resultados com até 85% de sensibilidade e acurácia mesmo tendo os experimentos sido executados sobre vetores de características compostos por medidas estatísticas muito simples. O uso de clusterização também se mostrou promissor. Em especial, o algoritmo x-means conseguiu se aproximar ao número de instâncias esperados por classe. Apesar de serem feitos vários testes com os algoritmos de extração de regras de associação, não foram obtidos relacionamentos de importância. Este fato pode ser devido a não terem sido incluídos os dados clínicos das pacientes, apenas terem sido usados dados numéricos nos atributos de característica das imagens. O software WEKA respondeu bem às necessidades de manipulação dos algoritmos de mineração e análise de resultados. Como prosseguimento deste trabalho pretende-se ampliar o banco de dados para que, além de armazenar imagens, tenham-se informações relevantes do paciente, tais como dados do histórico clínico e outras informações que possam permitir a obtenção de melhores regras de associação. Também se pretende testar outras características como as geoestatísticas e medidas fractais para melhorar o diagnóstico precoce de doenças da mama.

6. AGRADECIMENTOS

Essa pesquisa é apoiada pelo CNPQ – Conselho Nacional de Pesquisa e pela CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Ministério da Educação, Programa Pro - Engenharias PE021-2008 e Pro CAD-NF, nº 540/2009.

7. REFERÊNCIAS

Antonie, M. L.; Zaïane, O. R.; Coman, A. 2001. Application of Data Mining Techniques for Medical Image Classification. Proceeding of the 2nd international workshop in Multimedia, Data Mining, San Francisco, USA.

Araújo, A. V. 2006. Árvores de decisão fuzzy na mineração de imagens do sistema footScanAge. Dissertação de Mestrado. Universidade Federal do Paraná.

Araújo, A.V.; Bellon, O.R.P.; Silva L.; Vieira, E. V.; Cat, M. 2007. Aplicando Mineração de Imagens para Auxiliar na determinação da Idade Gestacional em Recém-Nascidos. Grupo Imago de Pesquisa em Visão Computacional, Universidade Federal do Paraná.

Arora, N.; Martins, D.; Ruggerio, D.; Tousimis, E.; Swistel, A.; Osborne, M. P. 2008. Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer. The American Journal of Surgery 196, pp. 523–526.

Borchartt. T. B.; Resmini, R.; Santos, A. M.; Conci, A.; Silva, A. C. 2011. Extração e Análise de Características em Termogramas para Auxilio ao Diagnóstico de Câncer de Mama. Congresso de Métodos Numéricos em Engenharia. Coimbra, Portugal.

Carvalho, L. A. V. 2001. Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração. 8. ed. São Paulo: Érica, p. 234.

Celinski, T.M.; Bellon, O.R.P. 1998. Métodos de agrupamento. Tutorial XI SIBPRAPI.

Deepa, S. D. 2011. Association Rule Mining Based on Image Content. International Journal of Information Technology and Knowledge Management.

Dheeba, J. 2010. Detection of Microcalcification Clusters in Mammograms using Neural Network. International Journal of Advanced Science and Technology.

Djeraba, C. 2001. Relationship extraction from large image databases. Proceedings of the 2nd International Workshop on Multimedia Data Mining, pp. 44-49, EUA.

Eddaoudi, F. 2011. Microcalcifications Detection in Mammographic Images Using Texture Coding. Applied Mathematical Sciences, Volume. 5, n. 8, 381 – 393.

Fayyad, U.; Djorgovski, S.G.; Weir, N. 1996 - a . Automating the analysis and cataloging of sky surveys. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp.473-493.

Fayyad, U.; Shapiro, G.; Smyth, P. 1996. Knowledge discovery and data mining: towards a unifying framework. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 1-34.

Ferrero, G.A. 2005. Detecção de Patrões em imagens médicas, Dissertação de Mestrado, Universidade Politécnica de Madrid (em Espanhol).

Gonzáles, R. C.; Woods, R. E. 2000. Processamento de imagens digitais. Edgard Blücher.

Hall, M.; Frank E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, n.1.

Hipp, J.; Güntzer, U. Nakhaeizadeh, G. 2000. Algorithms for association rule mining-a general survey and comparison. ACM SIGKDD Explorations, v.2, n.1, pp.58-63.

INCA - Instituto Nacional do Câncer. Disponível em: http://www2.inca.gov.br, último acesso em: 22-02-2012.

Ishwar, K. S. 2001. Mining Association Rules between Low-level Image Features and High-level Concepts, Department of Computer Science and Engineering, Oakland University, Rochester.

Kopersi, K.; Marchisio, G. B. 2000. Multi-level indexing and GIS enhanced learning for satellite imageries. Proceedings of the First International Workshop on Multimedia Data Mining. EUA.

Landwehr, N.; Hall, M. M.; Frank, E. 2003. Logistic Model Trees. 14th European Conference on Machine Learning, pp. 241-252.

Maimon, O.; Rokach, L. 2010. Data Mining and Knowledge Discovery Handbook. Second Edition.

Mert, A. 2011. Breast Cancer Classification by Using Support Vector Machines with Reduced Dimension. Dept. of Navigation Eng. Piri Reis University.

Motta, L. 2010. Obtenção automática da região de interesse em termogramas frontais da mama para o auxílio à detecção precoce de doenças. Dissertação de Mestrado, Universidade Federal Fluminense.

Narasimhan, H. 2009. Contribution-Based Clustering Algorithm for Content-Based Image Retrieval. College of Engineering, Guindy Anna University Chennai.

Ordonez, C.; Omiecinski, E. 1999. Discovering association rules based on image content. Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries Conference, pp. 38-49, EUA.

Proeng, Processamento e analise de imagens aplicadas à mastologia, disponível em http://visual.ic.uff.br/proeng, último acesso em: 22-02-2012.

Quinlan, R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann. San Mateo, CA.

Rajendran, P. 2010. Novel Fuzzy Association Rule Image Mining Algorithm for Medical Decision Support System. International Journal of Computer Applications.

Resmini, R. 2011. Análise de imagens térmicas de mama usando descritores de textura. Dissertação de Mestrado, Universidade Federal Fluminense.

Rolando, R. 2006. Evolutionary Neural Networks Applied To The Classification Of Microcalcification Clusters In Digital Mammograms. IEEE Congress on Evolutionary Computation.

Samma, A.S.B. 2009. Adaptation of K-Means Algorithm for Image Segmentation. World Academy of Science, Engineering and Technology 50.

Serrano, R. C. 2010. Análise da viabilidade do uso do coeficiente de Hurst e da lacunaridade no auxílio ao diagnóstico precoce de patologias da mama. Dissertação de Mestrado, Universidade Federal Fluminense.

Siddheswar, R.; Rose H. T. 1999. Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. School of Computer Science and Software Engineering Monash University.

Software ImageJ, disponível em: http://rsbweb.nih.gov/ij/, último acesso em: 22-02-2012.

Soh, L.; Tsatsoulis, C. 1999. Segmentation of satellite imagery of natural scenes using data mining. IEEE Transactions on Geoscience and Remote Sensing, v.37, n.2. pp. 1086-1099.

Vailaya, A.; Zhong, Y.; Jain, A.K. 1996. A hierarchical system for efficient image retrieval. Proceedings of the 13th, Conference on Pattern Recognition. Áustria.

Vieira, E.V.; Bellon, O. R. P.; Silva, L. 2002. Mineração de imagens. Revista de Informática Teórica e Aplicada. Porto Alegre - RS, n. 2, pp. 67-96.

Wang, Q.; Chung, S.M. 2001. Content-based retrieval and data mining of a skin cancer image database. Proceedings o the International Conference on Information Technology: Coding and Comp, pp.611-615. EUA.

Witten, I.H.; Frank, E. 2000. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann.

Yi, H.; Rajan, D.; Chia, L. T. 2005. ARIRS :Association Rule Based Image Retrieval System, International Workshop for Advanced Imaging Technology (IWAIT'05).

Zaïane, O. R. 2002. Mammography Classification by an Association Rule-based Classifier. International Workshop on Multimedia Data Mining.

Zaïane, O.R.; Han, J. 2000. Discovering Spatial Associations. Images Data Mining and Knowledge Discovery: Theory, Tools, and Technology II, SPIE 14th International Symposium on Aerospace/Defense Sensing, Simulation and Controls. EUA.

Zhang, J.; Hsu, W.; Lee, M. L. 2001. Image mining: issues, frameworks and techniques. Proceedings of the 2nd International Workshop on Multimedia Data Mining, pp. 13-20.

8. DIREITOS AUTORAIS