Análise Comparativa de Métodos de Aprendizagem de Máquina para Classificação de Massas em Mamografias

Matheus Cordeiro de Melo[1], Andy Anand Gajadhar[1], Leonardo Vidal Batista[1]

[1]Centro de Informática - Universidade Federal da Paraíba (UFPB) – João Pessoa - PB - Brasil

{matheus.melo, andygajadhar, leonardo}@ci.ufpb.br

Abstract*. The breast cancer is a common cancer among women and mammography is important for early detection of the disease. This work presents a comparative analysis of methods for classification, as malignant or benign, of masses found in mammographic images. The methodology consists in building a test database, mass segmentation, attributes extraction and the classification through machine learning algorithms, using the* Weka *tool. Tests show that the best CCR was 85.09% achieved with Multilayer Perceptron algorithm.*

Resumo. *O câncer de mama é um tipo de câncer comum entre as mulheres, e o exame mamográfico é importante para detectar a doença em fase inicial. O presente trabalho apresenta uma análise comparativa da classificação, em malignos ou benignos, de massas encontradas em imagens mamográficas. A metodologia consiste na, construção de uma base de testes, segmentação de massas, extração de atributos e classificação a partir de algoritmos de aprendizagem de máquina, utilizando a ferramenta* Weka*. Testes resultaram em uma taxa de acerto de 85,09%, atingida com o algoritmo Multilayer Perceptron.*

1. Introdução

O câncer de mama é o segundo tipo de câncer mais frequente no mundo e o mais comum entre as mulheres, respondendo por 22% dos novos casos a cada ano [INCA 2014]. Para detectar precocemente esse câncer, é muito utilizada a mamografia. No contexto dos Sistemas de Auxílio ao Diagnóstico médico (Computer-Aided Diagnosis - CAD), é de interesse segmentar os achados mamográficos com o intuito de separá-los do restante da mama, extrair características relevantes e aplicar técnicas de aprendizagem de máquina que forneçam uma sugestão diagnóstica, ou uma segunda opinião, a fim de auxiliar o médico a detectar e classificar eventuais nódulos.

Vários métodos têm sido desenvolvidos com o intuito de auxiliar a detecção precoce do câncer de mama. Em [Shah et. al. 2014] é descrito um sistema que aumenta a detecção de um possível tumor na mamografia a partir de algumas técnicas de pré-processamento. Em [Duarte et. al. 2013] foi desenvolvido um CAD para classificação de lesões mamográficas, extraindo características das imagens através do operador de Padrão Binário Completo (CLBP) e classificando a partir de uma rede neural artificial de Função de Base Radial (RBF).

O presente trabalho apresenta uma análise comparativa da eficiência de algoritmos de aprendizagem de máquina em classificar, como malignos ou benignos, massas encontradas em imagens mamográficas. Para realizar uma classificação precisa, foi necessário utilizar uma técnica de segmentação de massas eficaz e extrair atributos relevantes.

2. Materiais e Métodos

Foi utilizado o banco de imagens *Digital Database for Screening Mammography* (DDSM). O DDSM contém aproximadamente 2600 casos, em que cada caso possui duas imagens de cada mama (correspondentes às incidências craniocaudal e médio-lateral oblíqua), juntamente com informações sobre o paciente, a imagem e o diagnóstico, confirmado por exame anátomo-patológico, aqui considerado como padrão-ouro para avaliação dos classificadores.

Segmentação, quando utilizada em imagens mamográficas, é um processo complicado, devido em grande parte à anatomia da mama, ao baixo contraste e à não homogeneidade da exposição [Martí et. al. 2007].

Neste trabalho, foi implementado o algoritmo proposto em [Barbosa Filho 2010], baseado na técnica de Crescimento de Regiões e em Árvore de Decisão, voltado especificamente para ser utilizado na segmentação das massas durante o pré-processamento das mamografias e integrado em forma de *plug-in* no software de processamento de imagens ImageJ. Parte da segmentação utilizando o ImageJ é mostrada na Figura 1.

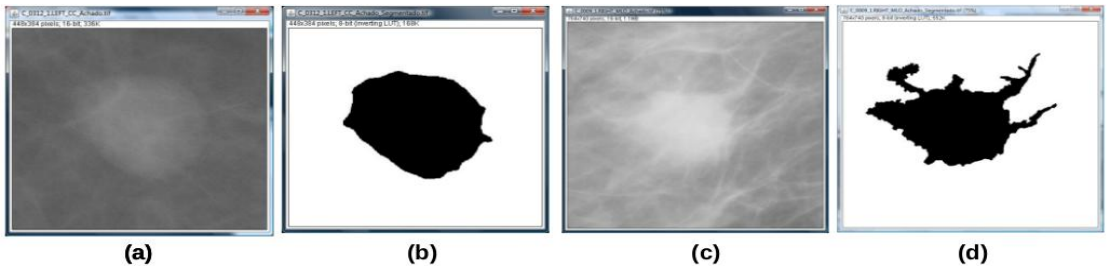

Figura 1. Segmentação de achado em mamografia utilizando o ImageJ. (a) Achado benigno; (b) Imagem segmentada de (a); (c) Achado maligno; (d) Imagem segmentada de (c).

O método proposto neste trabalho é mostrado, em linhas gerais, na Figura 2. Na segmentação, as regiões de interesse das mamografias do DDSM selecionadas por especialistas são segmentadas a partir do *plug-in* do ImageJ, citado anteriormente.

Por se tratar de um trabalho em andamento e os atributos de forma constituírem uma importante fonte de dados para o processo de classificação de achados mamográficos [Menechelli et. al. 2010], na etapa de extração de atributos são extraídos apenas atributos de forma das imagens segmentadas através da ferramenta Matlab. Dessa maneira, e seguindo propostas anteriores [Poel et. al. 2007], o vetor de características de cada imagem foi formado pela solidez, excentricidade, circularidade e área convexa da massa segmentada.

Depois de extrair os atributos de forma, uma base de características é construída para que seja utilizada na etapa de classificação. Nessa última etapa, foi utilizada a ferramenta Weka, a qual fornece implementações de diversos algoritmos de aprendizagem de máquina [Witten e Frank 2005]. Os algoritmos utilizados foram o *Multilayer Perceptron*, o *RBF Network*, o *Naive Bayes*, o *KNN*, o *Random Forest* e o *FT*, por serem muito utilizados na literatura da área. Para cada algoritmo foi aplicado sobre os dados o método de validação cruzada com 10 grupos, com o objetivo de classificar os achados em malignos ou benignos. Para avaliar os classificadores, foram adotadas as seguintes métricas: taxa de acerto, sensibilidade e especificidade [Zhu et. al. 2013].

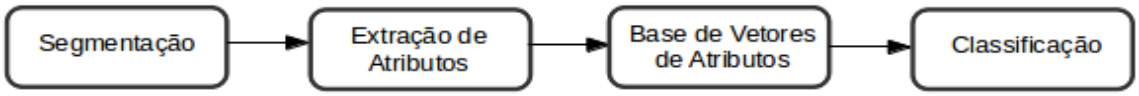

Figura 2. Esquema da sequência de etapas realizadas

3. Resultados

Para a realização dos testes foram utilizadas 161 imagens mamográficas selecionadas aleatoriamente do banco DDSM, onde 103 são benignas e 58 são malignas. Sobre essas imagens são executadas as etapas ilustradas na Figura 2. Deve-se destacar que as imagens de treinamento não são incluídas nos testes de classificação. A Tabela 1 mostra o resultado obtido com cada algoritmo utilizado na etapa de classificação. Observa-se que o melhor classificador, no que diz respeito à taxa de acerto e à sensibilidade, foi o *Multilayer Perceptron* com parâmetros default do Weka, atingindo 85,09% e 77,59%, respectivamente.

Tabela 1. Resultado da classificação por cada algoritmo

Algoritmos	Taxa de acerto %	Sensibilidade %	Especificidade %
Multilayer Perceptron	85,09	77,59	89,32
RBF Network	83,85	72,41	90,29
Naive Bayes	80,12	75,86	82,52
KNN (k = 5)	80,75	62,01	91,26
Random Forest	81,37	74,14	85,44
FT	80,75	72,41	85,44

4. Conclusão

O advento de novas tecnologias está mudando drasticamente o modo de interpretação de imagens médicas. Sistemas de CAD e de Recuperação de Imagens por Conteúdo (CBIR) auxiliam especialistas na identificação e classificação de possíveis nódulos. Uma segmentação adequada propicia uma extração de atributos eficaz e consequentemente uma classificação também eficaz. Devido à diferença entre as bases de testes, ainda não é possível uma comparação direta totalmente justa deste trabalho com o apresentado em

[Duarte et. al. 2013], mas aqui já são apresentados resultados competitivos, como mostrado na Tabela 1.

Como observado na seção 3, foram extraídos apenas atributos de forma das imagens segmentadas, apresentando resultados promissores. Como investigações futuras, pretende-se utilizar mais atributos de forma e contorno, além de atributos de textura e brilho, e aplicar métodos de seleção de atributos como Distância entre Classes e Seleção Sequencial para Frente utilizados em [Zhu et. al. 2013], visando validar o método proposto e integrá-lo a um sistema CAD, já em desenvolvimento no âmbito do grupo de pesquisa.

5. Referências

INCA (Instituto Nacional do Câncer) (2014). Disponível em: http://www.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama. Acessado em: 28 de Março de 2014.

Shah, N. N.; Ratanpara, T. V.; Bhensdadia, C. K. Early Breast Cancer Tumor Detection on Mammogram Images. International Journal of Computer Applications, v. 87, n. 14, pp. 14-18, 2014.

Duarte, Y. A. S.; Nascimento, M. Z.; Oliveiras, D. L. L. Algoritmo de Extração de Textura Baseado em Wavelet e CLBP para Classificação de Lesões em Mamogramas. XIII Workshop de Informática Médica, pp. 174-183, Maceió, Brasil, 2013.

Martí, R.; Oliver, A.; Raba, D.; Freixenet, J. Breast Skin-Line Segmentation Using Contour Growing. 3rd Iberian Conference on Pattern Recognition and Image Analysis, pp. 564-571, Girona, Spain: Springer-Verlag, 2007.

Barbosa Filho, J. R. B. Segmentação Automática de Massas Mamográficas Através do Crescimento de Regiões e Árvore de Decisão. 2010. 27 f. Monografia (Bacharelado em Ciência da Computação) – Departamento de Informática, Universidade Federal da Paraíba, João Pessoa. 2010.

Menechelli, R. C.; Ribeiro, P. B.; Schiabel, H. Desenvolvimento de Um Software Classificador da Forma de Nódulos Mamográficos Segmentados Utilizando A Rede Neural Artificial Multilayer Perceptron(MLP). VI Workshop de Visão Computacional. Presidente Prudente, Brasil, 2010.

Poel, J. K. D.; Mascena, E. N.; Pires, G. M.; Honório, T. C. S.; Medeiros, T. F. L.; Batista, L. V. Um Sistema Para Diagnóstico Auxiliado por Computador Voltado Para Imagens Mamográficas: Desempenho da Busca Baseada em Conteúdo na Recuperação de Achados. VII Workshop de Informática Médica. Porto de Galinhas, Brasil, 2007.

Witten, I. H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. 2 ed. San Francisco: Elsevier, 2005. Cap. 9, p. 366.

Zhu, M.; Xu, C.; Yu, J.; Wu, Y.; Li, C.; Zhang, M.; Jin, Z.; Li, Z. Differentiation of Pancreatic Cancer and Chronic Pancreatitis Using Computer-Aided Diagnosis of Endoscopic Ultrasound (EUS) Images: A Diagnostic Test. Plos One, v. 8, 2013.