ANÁLISE DE ATRIBUTOS DE INTENSIDADE E TEXTURA NA CLASSIFICAÇÃO DE DENSIDADE MAMÁRIA

P. C. Carneiro* e A. C. Patrocinio*

*Engenharia Biomédica, Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia, Brasil

e-mail: pedrocarneiro@ebio.ufu.br

Resumo: A densidade mamária está fortemente associada com ao risco de se contrair câncer de mama. Dessa forma, este trabalho tem por objetivo investigar atributos de imagens extraídos de histograma e de descritores de textura, a fim de separar imagens mamográficas por grau de densidade mamária, através de técnica de *clustering*. Foram utilizadas 307 imagens mamográficas coletadas do banco digital *INbreast*. Após a seleção dos atributos dos dados normalizados, foi obtido um índice de 85,01% de acerto na classificação das imagens dentro das quatro classes, quando os descritores de textura "energia", "variância" e "correlação" foram utilizados simultaneamente junto à técnica de *clustering.*
Palavras-chave: *clustering,* densidade mamária, extração de atributos, mamografia.

Abstract: *Mammographic density is well-known for being an important indicator of the risk of breast cancer. This way, this paper aims to investigate image features based on histogram and texture descriptors in order to separate mammographic images into the four patterns of breast density, through a clustering technique. 307 mammographic images were used from the INbreast database. After the feature selection of the normalized data, there was 85,01% accuracy on the mammogram classification in the four classes, whenever the following texture descriptors "energy", "variance" and "correlation" were used simultaneously with the clustering technique.*
Keywords: *clustering, breast density, feature extraction, mammography.*

Introdução

O *Breast Imaging Reporting and Data System* (BI-RADS™), proposto pelo Colégio Americano de Radiologia (ACR), é o nome dado ao sistema que visa padronizar e uniformizar os laudos de mamografia entre médicos e especialistas da área [1]. Assim quatro padrões foram criados para classificar a mama a partir da densidade mamária, sendo eles:

• Padrão 1: a mama é predominantemente adiposa (gordurosa);

• Padrão 2: a mama é parcialmente adiposa (com densidades de tecido fibroglandular ocupando de 26% a 50% do volume da mama);

• Padrão 3: a mama é densa e heterogênea (51% a 75% de tecido fibroglandular);

• Padrão 4: a mama é muito densa, apresentando mais de 75% de tecido fibroglandular.

A importância do estudo da densidade mamária se deve ao fato da direta relação entre o risco de se desenvolver e se contrair câncer de mama e o tipo de tecido predominante nesse órgão [2-5]. Mulheres que possuem acima de 75% de tecido fibroglandular na mama (mama densa) têm de quatro a cinco vezes mais risco de se contrair câncer em comparação com mulheres de mama pouco densas [6].

A abordagem desse trabalho é de que mamogramas classificados por diferentes padrões de densidade mamária sejam representados por tecidos diferentes, e consequentemente com diferentes características, isto é, cada padrão deverá apresentar valores distintos para os atributos analisados [6].

Como a diferença de intensidade de *pixel* entre imagens dos diferentes padrões é significante, a utilização de atributos baseados em medidas de intensidade de tons de cinza extraídas do histograma e atributos de textura como descritores de Haralick vem sendo constantemente estudados [3, 6-8], além de outros trabalhos que utilizam-se de atributos para quantificar os objetos nas imagens [9-12].

O objetivo deste trabalho é extrair atributos de imagens mamográficas a partir de atributos de intensidade e descritores de textura de Haralick, separando-as nas quatro classes de densidade mamária do BI-RADS™, utilizando uma técnica de *clustering.*

Materiais e métodos

Foram utilizadas 307 imagens do banco digital *INbreast Database* [13], já laudadas dentro dos quatro padrões de densidade mamária. Destes 307 mamogramas, 103 pertencem ao Padrão 1 (P1 - adiposa), 104 ao Padrão 2 (P2 – parcialmente adiposa, 26 a 50% de tecido fibroglandular), 73 imagens ao Padrão 3 (P3 – denso, 51 a 75% de tecido fibroglandular), e as demais 27 caracterizadas como Padrão 4 (P4 – muito denso).

As imagens salvas em formato DICOM (*Digital Imaging and Communications in Medicine*), de 12 bits por *pixel*, possuem tamanho de 3328 x 4084 ou 2560 x 3328 *pixels* dependendo do tamanho da mama da

paciente. Elas foram obtidas no mesmo equipamento, *MammoNovation Siemens FFDM*, estando sob vista médio lateral oblíqua (MLO) e crânio-caudal (CC).

A textura em uma imagem ou objeto contém informações sobre a distribuição espacial de variações de intensidade dentro de uma faixa de valores, enquanto a intensidade representa apenas o nível de cinza da imagem [14,15].

Os descritores de textura de Haralick [10] utiliza da matriz de co-ocorrência de níveis de cinza (SGLD – *Spatial Gray-Level Dependence*) para calcular a probabilidade de ocorrência combinada de direção e distância entre pares de *pixel* com valores de intensidade semelhantes, separados por uma distância "d", em quatro ângulos "θ", sendo eles 0º, 45º, 90º, 135º.

As matrizes de co-ocorrência levam em consideração a relação entre dois *pixels* por vez, sendo o primeiro chamado de *pixel* de referência e o segundo chamado de *pixel* vizinho.

Foram implementados tanto atributos extraídos de histograma quando descritores de textura de Haralick.

Os atributos extraídos do histograma foram: média de intensidade de pixels; valor de intensidade do maior pico do histograma (moda); menor intensidade do histograma; maior intensidade do histograma; porcentagem da maior intensidade em relação à intensidade máxima da escala de níveis de cinza; diferença da média para o menor valor; módulo da diferença da média para o maior valor; quantidade de pixels maiores que o pico do histograma; gradiente (maior intensidade subtraído da menor intensidade).

Já para os descritores de Haralick, foram implementados um conjunto de 13 atributos, sendo eles: energia ou uniformidade; contraste; correlação; variância; momento da diferença inversa; média da soma; variância da soma; entropia da soma; entropia; variância da diferença; entropia da diferença; medida de informação de correlação 1; medida de informação de correlação 2.

Cada imagem gerou um valor numérico para cada um dos atributos. Tendo esses resultados, foi feita uma média aritmética simples e o desvio padrão para atributos de imagens de mesma classe, podendo comparar os valores obtidos para cada um dos padrões de densidade.

Como cada atributo para cada classe possui um valor, inicia-se a etapa de seleção dos atributos, isto é, selecionar qual ou quais atributos podem melhor diferenciar as imagens dentro das quatro classes.

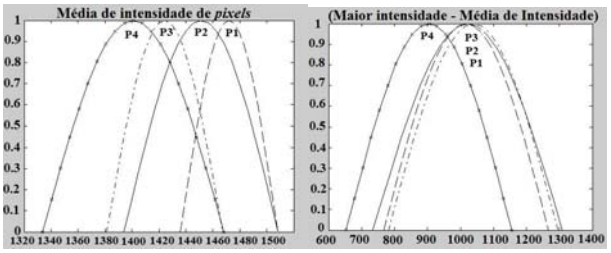
A seleção foi feita baseando-se na Distribuição Normal (Gaussiana). Esse método é inteiramente descrito por parâmetros de média e desvio padrão, ou seja, conhecendo-se estes valores é possível determinar qualquer probabilidade em uma distribuição normal. Quanto menor a sobreposição das curvas gaussianas entre as quatro classes, mais significativo deve ser o atributo.

Após a etapa de extração e seleção dos atributos, foi proposta a técnica de *clustering K-Means* [16]. Esse método objetiva particionar 'n' observações dentre 'k' clusters, onde cada observação pertence ao cluster mais próximo de sua média.

O *K-Means* é uma heurística de agrupamento não hierárquico que busca minimizar a distância dos elementos a um conjunto de k centros dado por $\chi=\{x1,x2,...,xk\}$ de forma iterativa. A distância entre um ponto pi e um conjunto de clusters, dada por $d(pi,\chi)$, é definida como sendo a distância do ponto ao centro mais próximo dele.

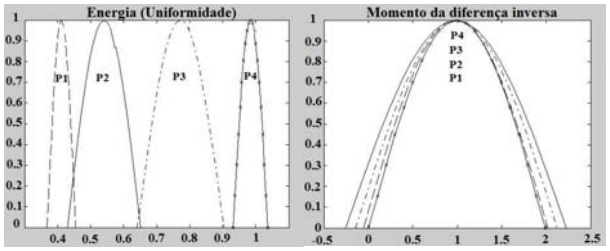
Resultados

Nas Figuras 1 e 2 são mostradas as curvas da distribuição normal de dois atributos de intensidade e dois atributos de Haralick, respectivamente.



(a) (b)

Figura 1: Distribuição Normal de atributos do histograma – (a) Média de intensidade de pixel com pouca sobreposição e (b) 'Maior intensidade – Média de intensidade' com muita sobreposição.



(a) (b)

Figura 2: Distribuição Normal de descritores de Haralick – (a) Energia e (b) Momento da diferença inversa.

Para os atributos de intensidade, o atributo de menor sobreposição foi o apresentado na Figura 1a. Todos os demais para esse tipo de característica apresentaram muitas sobreposição entre os padrões, a exemplo do mostrado na Figura 1b, e por isso não foram utilizados com o método de *clustering K-Means.*

Já para os atributos de textura (Haralick), o atributo Energia (Figura 2a) apresenta-se com pouca sobreposição caracterizando boa separabilidade enquanto, a exemplo do atributo da Figura 2b, as curvas se apresentam muito sobrepostas.

Após a análise de todas as curvas de distribuição normal, foi observado que os melhores atributos são: "energia", "correlação", "variância" e "média da soma". Esses descritores de Haralick foram inseridos no método de *clustering K-Means* e a porcentagem de acerto na classificação é apresentada na Tabela 1.

Tabela 1: Porcentagem de acerto do classificador para os atributos analisados

Atributo	Porcentagem de acerto
Energia	75,57%
Energia e Correlação	76,87%
Energia, Correlação, Variância e Média da Soma	79,8%
Energia e Variância	82,08%
Energia, Variância e Correlação	85,01%

Além do índice de acerto, foi quantificado também o número de inversões de classificação, ou seja, quantas imagens deveriam ser atribuídas a um padrão, mas durante a técnica de *clustering* foram agrupadas a uma classe diferente da qual ela realmente pertencia. Esses resultados são mostrados na Tabela 2, na qual a porcentagem de erros é com relação ao número total de erros observado.

Tabela 2: Erros da classificação em cada padrão.

Padrão correto ➔ Padrão classificado	Porcentagem de erros
P1 ➔ P2	5,55%
P2 ➔ P1	4,2%
P1 ➔ P3	3,56%
P3 ➔ P1	3,23%
P2 ➔ P3	28,15%
P3 ➔ P2	33,91%
P2 ➔ P4	3,9%
P4 ➔ P2	2,6%
P3 ➔ P4	8,1%
P4 ➔ P3	6,8%

Discussão

A distribuição normal (gaussiana) do atributo "energia", mostrado na Figura 2a, apresentou a menor sobreposição dentre todos os atributos de Haralick testados. Analisando a distribuição normal apenas do atributo de intensidade (Figura 1a), a "média de intensidade de *pixels*" foi a que apresentou a menor sobreposição dentre todos os atributos de intensidade. Entretanto o atributo de intensidade "módulo da diferença da média para o maior valor" (Figura 1b) e o descritor de textura "momento da diferença inversa" (Figura 2b) apresentaram curvas sobrepostas para os quatro padrões, impossibilitando a diferenciação.

Após a extração de atributos notou-se que os atributos de intensidade, em geral, apresentaram mais sobreposição entre as classes que os atributos de textura (descritores de Haralick). Esses atributos de intensidade apresentaram alta variância dentro da mesma classe, o que pode ser explicado pela presença de grandes lesões nodulares nas imagens, alterando o nível de intensidade de *pixel* entre elas.

Já para os descritores de textura, os resultados foram melhores. O atributo "energia" apresentou valores mais altos para o padrão de densidade 4, indicando maior uniformidade dessa classe dentre as demais.

O atributo "correlação" é um indicador de uma estrutura implícita na textura ou um fundo suave, apresentando maiores valores para o padrão de densidade 1. O atributo "variância" está relacionado à variabilidade de intensidade de pixels da imagem enquanto que "média da soma" representa a média dos tons de fundo da imagem.

As distribuições normais de cada atributo serviram para selecionar as melhores características, para serem usados na técnica *K-Means*.

A partir da Tabela 1 verifica que os melhores resultados foram obtidos quando mais de um atributo é utilizado simultaneamente na técnica de *clustering.* No entanto, à medida que se adiciona um atributo não necessariamente o agrupamento tende a melhorar, tanto é que o índice de acerto com três e dois atributos simultâneos foi maior do que utilizando quatro atributos, obtendo índices de acerto de 85% e 82%, respectivamente.

Quanto às inversões cometidas, já era de se esperar que os padrões de densidade 2 e 3 fossem aqueles em que mais gerariam confusão. Isso ocorre devido à similaridade dos tipos de tecidos presentes nas mamas dessas classes e por serem classes intermediárias, dificultando a caracterização e alocação correta de casos. Vale ressaltar que não ocorreu inversão do padrão 1 para o padrão 4 e vice-versa, uma vez que são imagens bem distintas quanto ao tipo de tecido predominante.

Contudo deve-se considerar que as imagens são laudadas por médicos e um alto grau de subjetividade está intrínseco ao processo de quantificação de densidade por avaliação visual, portanto qualitativa, que é justamente o processo utilizado na avaliação por especialistas.

Conclusão

A classificação de imagens mamográficas por padrão de densidade é uma tarefa cada vez mais difícil e sujeito a resultados com alto grau de confusão devido à subjetividade intrínseca desse processo.

Com a utilização de técnicas de *clustering* torna-se possível automatizar essa tarefa com taxas de erro aceitáveis, desde que a extração de atributos seja adequada às características das imagens que envolvem o problema. Neste caso, a variação de intensidade de pixel e consequentemente a textura da imagem, está diretamente relacionada ao padrão de densidade mamária.

Neste trabalho mostrou-se que 85,01% dos mamogramas foram agrupados corretamente no seu padrão de densidade mamária com a utilização de descritores de textura. Quando mais de um atributo é

utilizado simultaneamente na técnica, o agrupamento se torna melhor, elevando a acurácia do método.

A próxima etapa será uma avaliação estatística da composição dos tipos de tecido dos mamogramas, a fim de avaliar as porcentagens de tecidos fibroglandulares descritas para cada padrão de densidade.

Agradecimentos

Agradecemos ao *Breast Research Group*, do INESC Porto, Portugal, pela disponibilização do banco de imagens, e a FAPEMIG pelo apoio financeiro.

Referências

[1] D'orsi C, Basset L, Feig S. Illustrated breast imaging reporting and data system American College of Radiology. Reston, Va, 1998.

[2] Zonderland HM et al. Diagnosis of Breast Cancer: Contribution of US as an Adjunct to Mammography 1. Radiology. 1999; 213(2):413-22.

[3]Petroudi S, Kadir T, Brady M. Automatic classification of mammographic parenchymal patterns: A statistical approach. In: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2003 Sep 17-21; Cancún, México. 2003. p. 798-801

[4]Cuzick J, Warwick J, Pinney E et al. Tamoxifen and breast density in women at increased risk of breast cancer. Journal of National Cancer Institute. 2004; 96:621-28.

[5] Mccormack VA, Dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiology Biomarkers. 2006; 15:1159-69.

[6] Oliver A, Freixenet J, Zwiggelaar R. Automatic classification of breast density. In: Image Processing, 2005. ICIP 2005. IEEE International Conference on. IEEE; 2005 Sep 14. p 1258-61.

[7] Muštra M, Grgić M, Delač K. Breast Density Classification Using Multiple Feature Selection. Automatika: Journal for Control, Measurement, Electronics, Computing and Communications. 2012; 53:362-72.

[8] Bosch A, Muoz X, Oliver A et al. (2006) Modeling and Classifying Breast Tissue Density in Mammograms. In: Computer Vision and Pattern Recognition. 2006; 1552-8.

[9] Sharma V, Singh, S. CFS–SMO based classification of breast density using multiple texture models. Medical & biological engineering & computing. 2014; 52(6): p. 521-9.

[10]Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. System Man Cybernetics. 1973; 610-21.

[11]Arivazhagan S, Ganesan L (2003) Texture classification using wavelet transform. Pattern Recognition Letters. 2003; 24:1513-21.

[12]Subashini T, Ramalingam V, Palanivel, S. Automated assessment of breast tissue density in digital mammograms. Computer Vision and Image Understanding. 2010; 114(1):33-43.

[13]Moreira, IC et al. INbreast: toward a full-field digital mammographic database. Academic radiology. 2012; 19(2): 236-48.

[14]Gonzalez RC, Woods RE. Processamento Digital de Imagens. 3a ed. Prentice Hall; 2010.

[15]Castleman, KR. Digital Image Processing. Curve and Surface Fitting; 1996. p.501-07.

[16]Hartigan J, Wong M. A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. 1979; 28:100-8.