DOI: 10.18605/2175-7275/cereus.v10n2p101-110

<<Recebido: 10/06/2018. Aceito: 12/07/2018>>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

Aplicação de Máquinas de Vetores de Suporte Na Classificação de Microcalcificações de Imagens Mamográficas

Application of Support Vector Machine in the Classification of Microcalifications of Mammographic Images

Fabiano Marcos de Lima[1], Thiago Alves de Souza[2], Robson Mariano da Silva[3]

[1] M.sC UFRRJ.
E-mail: fabianomarcos1@hotmail.com

[2] M.sC UFRRJ.

[3] D.sC UFRRJ.

RESUMO

O câncer de mama é um dos principais cânceres no mundo e é responsável por altas taxas de mortalidade, principalmente entre as mulheres. A mamografia é um método de obtenção de imagens indicados para tecidos mamários densos. Todavia, a imagem resultante apresenta ruídos e necessita de um especialista a fim de identificar a estrutura e a localização do tumor. O presente artigo utiliza Máquina de Vetor de Suporte na categorização de microcalcificação mamária em achados de exame de mamografia. O conjunto de dados utilizados contém informações referente a achados em exames de mamografias em 502 indivíduos. Os resultados obtidos evidenciam desempenho promissor do modelo proposto, visto que o mesmo no conjunto das 50 simulações realizadas obteve acurácia acima de 88,0%, sensibilidade superior a 86,0% e especificidade acima de 82,0%.

Palavras-chave: Inteligência Computacional. Máquina de Vetor de Suporte. Câncer de Mama.

ABSTRACT

Breast cancer is one of the leading cancers in the world and is responsible for high rates of mortality, especially in women. Mammography is an imaging method indicated for dense mammary tissues. However, the resulting image is noisy and requires a specialist to identify the structure and location of the tumor. The present article uses the Support Vector Machine in the categorization of mammary microcalcification in mammography examination findings. The data set contains information on mammography findings in 502 individuals. The obtained results show a promising performance of the proposed model, since the same in the set of 50 simulations obtained obtained accuracy above 88.0%, sensitivity higher than 86.0% and specificity above 82.0%.

Keywords: Computational Intelligence. Support Vector Machine. Breast Cancer

1. INTRODUÇÃO

O relatório de estimativas de incidências de Câncer no Brasil, realizado pelo Instituto Nacional do Câncer (INCA), mostra que em 2012, ocorreram 14,1 milhões de casos novos de Câncer e 8,2 milhões de óbitos em escala mundial (INCA,2018).

Segundo o mesmo relatório, houve um discreto predomínio de incidência quanto ao sexo masculino apresentou cerca de 53% e apresentando também aumento na taxa de mortalidade com cerca de 57% de óbitos. De modo geral, as maiores de taxas de incidências foram observadas em países desenvolvidos como América do Norte, Europa Ocidental, Japão, Coréia do Sul, Austrália e Nova Zelândia. Já as taxas intermediárias foram mais frequentes nas Américas Central e Sul, Leste Europeu e em grande parte do Sudeste Asiático (incluindo a China). As menores taxas foram vistas em grande parte na África e no Sul e Oeste da Ásia incluindo também a Índia. Nota-se que os tipos de Câncer que predominam nos países desenvolvidos são associados a fatores de urbanização e ao desenvolvimento.

Os Cânceres mais comuns nesses países são Pulmão, Próstata, Cólon, Reto e o de Mama feminina, ou seja, ocorrem com maior frequência nesses países (FERLAY *et al*.,2013). Dentre estes o artigo irá somente tratar no Câncer de Mama.

O Câncer de Mama é o segundo tipo de câncer mais comum entre as mulheres no Brasil e no mundo, respondendo por cerca de 28% dos casos novos a cada ano. Sua maior incidência ocorre em mulheres de 35 a 50 anos, e a cada 8 uma é diagnosticada com câncer de mama. Segundo dados do Instituto Nacional do Câncer, no ano de 2016 foram diagnosticados 57.960 novos casos de câncer de mama, com o total de óbitos de 14.388 decorrente de intercorrência dessa neoplasia (INCA, 2016). Desde o início das pesquisas sobre o câncer de mama, a melhor maneira para cura da doença é a detecção precoce (Mavroforakis, 2005). A detecção pode ser conseguida através do exame da mamografia, cujo método é o mais utilizado para o rastreamento do câncer de mama disponível hoje. A mamografia é uma forma particular de radiografia capaz de registrar imagens da mama com a finalidade de diagnosticar a presença ou ausência de estruturas que possam indicar a doença. Com esse tipo de exame pode-se detectar o tumor antes que ele se torne palpável. No entanto, a avaliação do exame de mamografia e o diagnóstico, que é realizado por um radiologista, requer bastante habilidade, porém há limitações na predição primária do câncer de mama.

Segundo Mavroforakis (2005), de 10% a 30% das mulheres que apresentam câncer de mama tiveram resultados negativos quando submetidas à mamografia, o que a crer que houve uma má interpretação dos exames. Distorções na interpretação e classificação de lesões por especialistas implicam um número maior de biópsias desnecessárias, ou seja, entre 65% a 85% das biópsias de mama são realizadas em lesões benignas. Com isso, há uma redução na relação custo benefício dos exames e, no pior caso, a não detecção da doença, caracterizando um diagnóstico falso-negativo.

Em virtude dos fatos mencionados, vários pesquisadores estão utilizando técnicas de inteligência computacional, no desenvolvimento de sistemas de apoio ao diagnóstico por computador (CAD), visando aumentar a taxa de detecção de Câncer. Dentre essas técnicas destacam-se as Redes Neurais Artificiais-RNAs (Azar & El-Said,2012; Tahmasbi *et a*l; 2011) e as Máquinas de Vetores de Suporte-SVMs (Menaka & Karpagavalli,2013; Huang *et al*,2017). Essas técnicas possuem a vantagem de serem robustas em um conjunto de dados ruidosos, apresentando um bom desempenho na análise de imagens.

Dessa forma, o objetivo desse trabalho é elaborar um modelo computacional estruturado em SVM, na classificação de malignidade de massa mamária obtidas em achados mamográficos.

As máquinas de Vetores de Suporte (SVM) é um método de aprendizagem supervisionada usado para estimar uma função que classifique dados de entrada em duas classes (Vapnik e Cortes,1998). São baseadas nos princípios de minimização do risco estrutural, que tem suas origens na teoria do aprendizado estatístico. O erro do algoritmo de aprendizagem juntos aos dados de validação (erro de generalização), é limitado pelo erro de treinamento mais um termo que depende da dimensão VC (dimensão Vapnik e Chervonenkis) (Semolini,2002), que é uma medida da capacidade de expressão de uma família de funções. O que se deseja é construir de hiperplanos tendo como estratégia a variação da dimensão VC, de modo que o risco empírico (erro de treinamento) e a dimensão VC sejam minimizados ao mesmo tempo.

Para que seja possível a SVM classificar as amostras que não são linearmente separáveis, necessita-se de uma transformação não linear que transforme o espaço de entrada em um novo espaço. A dimensão desse espaço deve ser suficientemente grande, e através dele, a amostra pode ser linearmente separável. Assim, o hiperplano de separação é definido como uma função de vetores retirados do novo espaço ao invés do espaço de entrada original. Segundo Haykin (2007), a construção depende do cálculo de um produto interno com uma função K de núcleo equação (1). A função K pode realizar o

mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

$$W(x) = \sum_{i=1}^{N} \alpha_1 - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (1)$$

As SVM possuem diferentes funções de núcleos que caracterizam seu modo de reconhecimento das padrões. Dentre as funções de núcleos mais utilizadas podemos citar: Linear, Polinomial (que manipula uma função polinomial, cujo grau pode ser definido durante os treinamentos), Gaussiano (corresponde a um espaço de características de dimensão infinita; a utilização desse núcleo permite que uma SVM apresente características de uma rede RBF) e Sigmoidal (permite que a SVM tenha comportamento semelhante ao de uma rede MLP).

A base radial é uma importante família de funções de núcleo, sendo muito utilizada em problemas de reconhecimentos de padrões não linearmente separáveis. Neste trabalho utiliza-se a base radial e está função é definida na equação (2). A correta definição do núcleo e de seus respectivos parâmetros possui forte influência nos resultados obtidos por uma SVM.

$$K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \quad (2)$$

2. MATERIAIS E MÉTODOS

O conjunto de dados utilizados consiste em 502 indivíduos, obtidos da base de dados pública do Instituto de Radiologia da Universidade Erlangen-Nuremberg, no período de 2003 a 2006. Contendo informações de achados em exames de mamografia (Birads, Margem, Densidade) e outras informações adicionais sobre o paciente.

A metodologia proposta é representada pela figura (1), onde o modelo computacional proposto é baseado em SVMs não linear, na classificação de malignidade em decorrência da existência de microcalcificações.

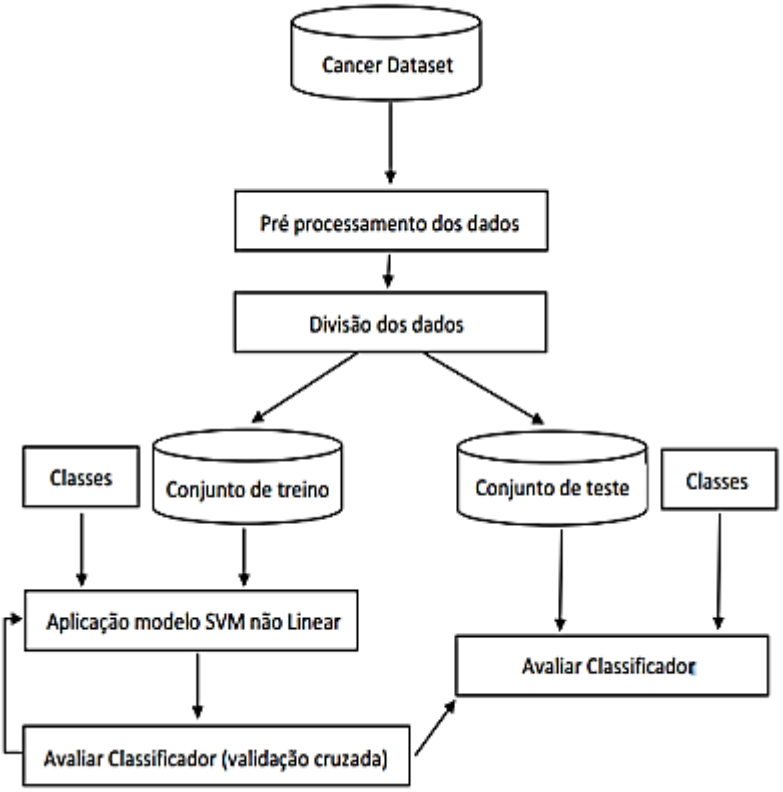

Figura1. Fluxograma da metodologia proposta

O pré processamento dos dados consistiu na normalização dos atributos e análise exploratória. Sendo amostra dividida ao mero acaso em dois grupos independentes, o conjunto de treinamento composto por 80% dos dados e o conjunto de teste com os 20% restantes. A função de *Kernel* utilizada e os valores dos parâmetros C e σ foram estimados através da avaliação dos resultados de simulações com amostras de treinamento utilizado validação cruzada. O modelo computacional proposto foi implementado utilizando o software R e o pacote *Kernlab* (Karatzoglou *et al*, 2014).

A performance do modelo SVM é avaliado pela acurácia de classificação ou precisão (ACC). A Acurácia é um grau de exame utilizado para avaliar o verdadeiro valor daquilo que está sendo medido, observado ou interpretado. (BONITA, 2010,p.4)

$$AAC = \frac{VP + VN}{VP + VN + FP + FN} \qquad (3)$$

Onde VP são os verdadeiros positivos, VN os verdadeiros negativos, FP falsos positivos e FN falsos positivos.

A Sensibilidade (S) é a medida utilizada em exames de diagnósticos que é a capacidade do teste de apresentar e de detectar os indivíduos que são verdadeiramente positivos, ou seja de diagnosticar corretamente os indivíduos que estão doentes. A Sensibilidade (S) pode ser definida como:

$$S = \frac{VP}{VP + FN} \tag{4}$$

Já a Especificidade (E) é a medida utilizada em exames de diagnósticos que é a capacidade do teste de apresentar e detectar os indivíduos que são verdadeiramente negativos, ou seja, de diagnosticar corretamente os indivíduos que estão não doentes. A Especificidade (E), pode ser definida como:

$$E = \frac{VN}{VN + FP} \tag{5}$$

Os parâmetros utilizados no modelo SVM não linear na classificação foram obtidos de forma experimental. A relação dos parâmetros está sumarizada na tabela 1.

Tabela 1. Parâmetros da SVM

Parâmetros	Valor
Simulações	50
Tipo do classificador	C-svc
Função de kernel	Rbf
Variância da função de kernel (σ)	[0,5; 0,25; 0,125; 0,05]
Parâmetro de regularização (C)	[0,1; 1,0; 10; 100]
Critério de parada (tol)	0,001

3. RESULTADOS

O modelo computacional SVM proposto nesse trabalho foi avaliado pela incorporação de todas as variáveis (Birads, Margem, Forma e Densidade) e excluindo uma variável por vez no conjunto de dados de pacientes portadores de microcalcificação mamárias. Nas tabelas (2,3,4,5 e 6) apresenta-se a comparação entre os resultados médio obtidos nas 50 simulações dos modelos (BFDM, BFM, BFD, BMD e MFD) com a aplicação do modelo computacional proposto.

Tabela 1. Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Margem, Forma e Densidade (BMFD).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	79,1±0,04	71,7±0,09	84,4±0,08	20,3±0,01	23,6±0,02
[Mediana±irq]	78,9±0,05	72,6±0,14	85,7±0,08	20,1±0,07	23,7±0,02
[Max-Min]	88,2-73,7	88,2-56,8	95,3-70,5	21,8-19,7	26,9-20,5
Melhor simulação	88,16	82,29	92,68	20,30	20,10

A simulação que apresentou melhor desempenho no conjunto de validação em termos da precisão total obteve valor de 88,16% na caracterização referente a malignidade de microcalcificação mamária. No que tange a especificidade, classificação referente a não malignidade da microcalcificação mamária o melhor valor obtido foi 92,68%.

Tabela 2. Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Forma e Densidade (BFD).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	80,2±0,04	69,9±0,07	89,6±0,08	20,5±0,01	22,5±0,02
[Mediana±irq]	80,3±0,05	68,4±0,08	90,6±0,08	20,6±0,01	22,1±0,01
[Max-Min]	88,2-75,0	82,4-56,8	100-68,9	22,2-20,1	29,0-20,1
Melhor simulação	88,20	77,10	97,60	22,00	22,70

Os resultados obtidos excluindo o modelo a variável Margem, mostram que o desempenho no conjunto de validação teve um aumento na performance na detecção da especificidade, ou seja, o modelo foi mais específico na detecção de pacientes portadores de microcalcificações benignas.

Tabela 4. Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Forma, Densidade e Margem (FDM).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	75,0±0,04	82,8±0,06	65,7±0,08	25,4±0,01	28,2±0,02
[Mediana±irq]	75,0±0,05	84,0±0,09	65,7±0,10	25,3±0,01	27,9±0,02
[Max-Min]	85,5-68,4	100-72,7	84,6-53,8	27,2-24,4	27,2-2,44
Melhor simulação	85,52	87,50	85,52	27,16	27,90

Os resultados obtidos excluindo do modelo a variável Birads, mostram que o desempenho no conjunto de validação teve uma uniformidade nos índices de acurácia, sensibilidade e especificidade

Tabela 5. Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Densidade e Margem (BDM).

	ACC(%)	S (%)	E(%)	Erro Treino	Erro VC
[Média ±sd]	79,3±0,03	70,4±0,08	87,7±0,06	20,7±0,01	27,7±0,02
[Mediana±irq]	78,9±0,04	70,6±0,09	88,3±0,08	20,6±0,01	22,7±0,02
[Max-Min]	85,5-73,7	89,3-55,5	95,5-78,0	22,0-19,9	26,8±20,3
Melhor simulação	85,52	74,30	95,12	22,01	23,00

Os resultados obtidos excluindo do modelo a variável Forma, mostram que o desempenho no conjunto de validação teve um aumento na performance na detecção da especificidade. Ato que nos possibilita conjecturar sobre a importância dessa variável na detecção de verdadeiros positivos (microcalcificações malignas).

Tabela 6. Resultado das médias de 50 simulações utilizando o modelo com as variáveis: Birads, Forma e Margem (BFM).

	ACC (%)	S (%)	E (%)	Erro Treino	Erro VC
[Média ±sd]	78,4±0,03	75,2±0,10	81,7±0,08	20,2±0,01	22,1±0,02
[Mediana±irq]	77,6±0,04	77,2±0,17	81,3±0,16	20,6±0,01	22,7±0,03
[Max-Min]	84,2-73,7	90,3-56,8	93,2-68,9	21,6-18,5	26,6-18,6
Melhor simulação	84,21	90,32	80,0	21,31	26,55

Os resultados obtidos excluindo a variável Densidade, considerando o modelo BFM, mostram que o desempenho no conjunto de validação foi significativo na performance na detecção da sensibilidade (S) acima de 90,00%.

4. DISCUSSÃO

Os altos índices e taxas de mortes causados pelo Câncer de Mama no Brasil e ao redor do mundo, justificam o desenvolvimento de pesquisas científicas voltadas para estratégias de auxilio no diagnóstico de detecção de doenças de tumores. Um diagnóstico mais precoce e mais preciso de doenças como o Câncer de Mama é crucial para um fator determinante para a eficácia e sucesso do tratamento. Dentro desse contexto, o presente

artigo apresenta os fundamentos das Máquinas de Vetores de Suporte não linear e propõe-se uma aplicação na classificação de Neoplasias Mamárias.

A exclusão no modelo da variável Densidade, mostrou que o desempenho no conjunto de validação teve um aumento na performance na detecção de verdadeiros positivos como negativos. Indicando que a variável Densidade é um fator que contribui para a piora do modelo na classificação das microcalcificações detectadas em exames de mamografia.

5. CONSIDERAÇÕES FINAIS

De acordo com a análise dos resultados foi possível evidenciar o desempenho promissor do modelo proposto, na caracterização de microcalcificação em dados mamográficos, visto que o mesmo no conjunto das 50 simulações realizadas obteve precisão total acima de 88%, sensibilidade superior a 86% e especificidade acima de 82%.

REFERÊNCIAS

AZAR, A.T.; El-Said, S.A., **Superior neuro-fuzzy classification systems.** Neural Computing and Applications, 23(1-supplement),55-72. (2012)

BONITA, R.,BEAGLEHOLE, R. e KJELLSTROM,T.,**Epidemiologia Básica 2° edição**. Livraria Santos Editora Com. Imp.Ltda.(2010)

BILSKA-WOLOAK, A.O.;FLOYD, C.E. JR.; LO, J.Y.; BAKER, J.A., **Computer aid for decision to biopsy breast masses on mammography**:validation on new cases. Academic Radiology, 12(6), 669-670.(2005)

FERLAY, J. et al. GLOBOCAN 2012 v1 .0**, cancer incidence and mortality worldwide. Lyon**, **France:** IARC, 2013. (IARC CancerBase, 11). Disponível em: <http://globocan.iarc.fr>. Acesso em: 14 set. 2013.

HUANG, M.W.;Chen, C.W.;LIN,W.C.;Ke,S.W.;TSAI,C.F. (2017), **SVM and SVM Ensembles in Breast CAncern Journal of Roentgenology**, 158(3), 521-526.

INCA (2016), Tipos de Câncer: Mama. Instituto Nacional de Câncer, URL: http://www2.inca.gov.br.br/wps/wcm/connect/tiposdecancer/site/home/mama. Acesso em 03/08/2017.

KARATZOGLOU, A.; SMOLA, A.; HORNIK, K. (2014), **Kernel-Based Machine Learming Lab Journal of Statistical Software**, 11(9).

MAVROFORAKIS, M.;GEORGIOU, H.;DIMITROPOULOS, N.; CAVOURAS,D.;THEODORIDIS, S., **Significance analysis of qualitative mammographic features**, using linear classifiers, neural networks and support vector machines. European Jounal of Radiology, 54(1), 80-89.(2005)

MENAKA,K.;KARPAGAVALLI,S.,BREAST **Cancer Classification using Support Vector Machine and Genetic Programming**. International Journal of Innivative Research in Computer and Communication Engineering, 1(7), 1410-1417.(2013)

SEMOLINI, R. , *"***Support vector machines, inferência transdutiva e o problema de classificação** *".* Dissertação de mestrado, UNICAMP, Campinas.(2002)

TAHMASBI, A.;SAKI,F.;SHOKOUHI, S.B., **Classification of benign and malignant masses based on Zernike moments**. Computer in Biology and Medicine. 41(8),726-735.(2011)

VAPNIK, C.; CORTES, V.N., **Support vector networks, Machine learnin Prediction.** PLOS ONE 12(1).(1995)