

Classificação de Tumores Mamários Utilizando Aprendizado de Máquina

R. Taveira*; R. Pires*

*Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, São Paulo, Brasil
E-mail: rictaveira@outlook.com

Introdução: O câncer de mama representa o tipo de maior incidência de câncer e a principal causa de morte por câncer em mulheres em todo o mundo. Em relatório feito para a Agência Internacional de Pesquisa sobre Câncer da Organização Mundial de Saúde (OMS), foram estimados cerca de 1,7 milhão de novos casos da doença (representando cerca de 25% de todos os cânceres em mulheres), além de 500.000 mortes (15% de todas as mortes por câncer em mulheres) ao redor do globo em 2012. O objetivo deste trabalho é criar um sistema que possa classificar automaticamente as imagens obtidas através de biópsias mamárias entre neoplasias malignas ou benignas.

Métodos: Um banco de imagens público contendo 7909 imagens de lâminas de biópsias de tecido mamário de 82 pacientes [1] está sendo utilizado para este trabalho. As imagens terão suas características extraídas e descritas pela técnica *Bag-of-Features*. A primeira fase da abordagem *Bag-of-Features* é a detecção de regiões de interesse de uma imagem baseada na representação de descritores. *Speeded-up Robust Features* é um descritor local de pontos de interesse que utiliza a matriz hessiana. Tais pontos de interesse são padrões locais representativos da imagem como cantos, *blobs* (que são regiões de uma imagem que compartilham propriedades em comum) e variações de contraste, e a partir deles são criados os vetores que representam diversas partes da imagem. Esses vetores são chamados de descritores de imagem. Neste sentido, cada região de interesse é representada por um ponto no espaço \mathbb{R}^n , para n características. Diferentes dicionários, contendo de 30 a 1000 palavras visuais serão definidos agrupando as características extraídas das imagens na fase de treinamento. Este trabalho propõe que 75% das imagens, escolhidas de maneira aleatória, sejam utilizadas na fase de treinamento e 25% na fase de teste. Máquinas de Vetores de Suporte serão utilizadas para a classificação. Vetores de entrada são mapeados de forma não linear em um espaço de características Z de alta dimensão. Neste espaço de características, um hiperplano ótimo é construído como superfície de decisão, de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima, baseado na teoria da minimização do risco estrutural [2].

Resultados: Os resultados ainda são preliminares, mas mostram-se promissores. Em teste com o grupo de imagens com ampliação do microscópio de 40x, que consiste de 625 imagens rotuladas por médico patologista como benignas e 1370 imagens rotuladas como malignas. 1000 palavras formaram o dicionário de palavras visuais. A classificação foi feita por Máquina de Vetores de Suporte com *Kernel* Linear. O resultado deste teste pode ser visto na Tabela 1.

Tabela 1 – Dados obtidos

		Classificação Proposta	
		Benigno	Maligno
Diagnóstico Médico	Benigno	0,82	0,18
	Maligno	0,22	0,78

Com base nesta matriz de confusão, é possível calcular as principais métricas de performance do sistema:

Acurácia = 80%; Sensibilidade = 82%; Especificidade = 78% e Precisão = 79%

Conclusão: A probabilidade de acerto da caracterização da neoplasia em torno de 80% (nos casos benignos e malignos) é considerada satisfatória para os resultados preliminares. Alterações no *Kernel* da Máquina de Vetores de Suporte e no número de palavras do dicionário podem promover melhorias no desempenho do sistema. Espera-se que o programa computacional desenvolvido para este trabalho possa auxiliar o patologista a ser mais produtivo e consistente nos diagnósticos.

Referências: [1] SPANHOL, Fabio A. et al. A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, v. 63, n. 7, p. 1455-1462, 2016

[2] CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. Machine learning, v. 20, n. 3, p. 273-297, 1995.