VI Encontro Acadêmico de Computação - UFMA
 ISSN: 2594-486X

Diagnóstico de Melanoma usando Função K de Ripley e Máquina de Vetores de Suporte

Lucas Bezerra Maia[1], Nigel da Silva Lima[1], Geraldo Braz Junior[1], João Dallyson Sousa de Almeida[1], Anselmo Cardoso Paiva[1]

[1]Núcleo de Computação Aplicada
Universidade Federal do Maranhão (UFMA)
Av. dos Portugueses, 1966 - Bacanga, São Luís - MA, 65080-805

{lucasmaia1202, nigelnaiguel.comp, ge.braz}@gmail.com

{jdallyson, anselmo.c.paiva}@gmail.com

Abstract. *Melanoma is the most lethal type of skin cancer compared to the others, however patients present high recovery rates when their illness is discovered in its primary stage. Several approaches to automatic detection and diagnosis have been explored by different authors, using pattern recognition and machine learning techniques. In this work, a model for automatic classification of melanoma skin cancer is proposed, through a supervised training process of support vector machines, using as characteristics, pieces of informations extracted with Ripley's K function. The results obtained are promising due to the sensitivity and precision rates found.*

Resumo. *O melanoma é o tipo de câncer de pele mais letal quando comparado com os demais, porém os pacientes apresentam índices de recuperação elevados se a doença for descoberta em sua fase inicial. Diversas abordagens de detecção e diagnóstico automático vêm sendo exploradas por diferentes autores, usando técnicas de reconhecimento de padrões e aprendizado de máquina. Neste trabalho, é proposto um modelo para classificação automática de câncer de pele melanoma, por meio de um processo de treinamento supervisionado de Máquina de Vetores de Suporte, tendo como características as informações extraídas com a função K de Ripley. Os resultados obtidos são promissores devido às taxas de sensibilidades e precisão encontradas.*

1. Introdução

O melanoma, mesmo sendo considerado o tipo de câncer de pele mais letal, apresenta índices de cura elevados quando diagnosticado em seu estágio inicial. A doença se manifesta em pele normal a partir do surgimento de uma pinta com tonalidade escura e bordas irregulares, ou de uma lesão pigmentada pré-existente, onde o tumor irá evoluir apresentando mudanças de cores e aumento da área da lesão [Soares 2008].

Uma das técnicas utilizadas para o diagnóstico de câncer de pele é a dermatoscopia, exame realizado com uma ferramenta que consiste em um amplificador, com uma fonte de luz não-polarizada, uma lente transparente e um líquido que fica entre o instrumento e a pele. O aparelho permite a avaliação de cores e microestruturas da epiderme, junção dermo-epidérmica e derme papilar que não são visíveis a olho nu.

Em [Nasr-Esfahani et al. 2016], foi proposto um modelo com técnicas de pré-processamento em uma base de 170 imagens dermatoscópicas para alimentar uma Rede Neural Convolucional (CNN) pré-treinada, chegando a uma acurácia de 81% como resultado. Com o mesmo objetivo, para uma base de 200 imagens, [Bakheet 2017] usam características de Histogramas de Gradientes Orientados (HOG) em conjunto com SVM para detecção de melanoma. O resultado apresentou 98,21%, 96,43% e 97,32% para sensibilidade, especificidade e acurácia, respectivamente.

O objetivo deste trabalho é propor uma metodologia de diagnóstico de câncer de pele por meio de um processo supervisionado de aprendizagem de máquina, usando Função K de Ripley [Ripley 1977] e Máquina de Vetores de Suporte [Soares 2008].

2. Fundamentação Teórica

2.1. Função K de Ripley

A maioria das estruturas no mundo natural não são homogêneas, mas exibem algum tipo de padrão espacial, seja ela em uma visão macro, como corpos celestes, ou micro, como um conjunto de células biológicas, sendo, portanto, possível ser resumido em um determinado padrão de pontos [de Oliveira Martins et al. 2007]. Esses padrões vêm sendo estudados por longo período na área da estatística e usam, geralmente, distâncias como medidas para objetos de estudo mapeados em um plano ou espaço [Haase 1995].

A Função K de Ripley é um método de análise comumente utilizada em dados espaciais. Essa função pode ser usada para resumir certos padrões de pontos, testar hipótese, estimar parâmetros e ajustar modelos [Braz Junior et al. 2014]. É definida por:

$$R(d,i) = \sqrt{\frac{Ak(i,j)}{N}},\ i \neq j \tag{1}$$

onde d é a distância entre o ponto a ser analisado e o ponto de referência, i, que é o centro do parâmetro área A, k é um função que verifica se o ponto j está dentro de A em relação a i e N é o número de pontos analisados.

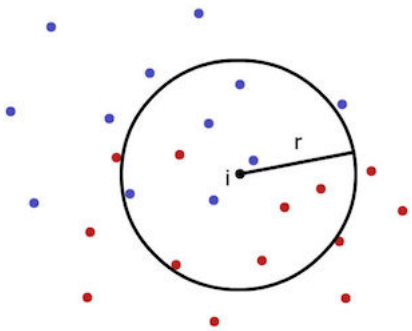

Figura 1. Análise de Ripley com área circular definida pelo raio r. Fonte: Autor

De uma forma geral, a função K de Ripley calcula uma relação do número de indivíduos de uma determinada espécie distribuída em uma região adotada, que pode ser circular, em anéis, quadrática ou até formas complexas.

2.2. Máquina de Vetores de Suporte

A máquina de vetores de suporte, do inglês *Support Vectors Machine - SVM*, é utilizada para tarefa de classificação e surgiram com o objetivo primário de categorizar padrões linearmente separáveis. O classificador SVM tem o trabalho de dividir duas classes através de um hiperplano que maximiza uma margem de separação denotada por δ. Na Figura 2, esse hiperplano está localizado entre as semi-retas H_1 e H_2, que separam as classes +1 e -1, respectivamente, atingindo seu ótimo quando a distância para as duas linhas for máxima. Dois pontos da classe +1 definem a reta H_1 e três, da classe -1, definem a reta H_2. Esses pontos são chamados de vetores de suporte (*Support Vector - SV*) [Soares 2008].

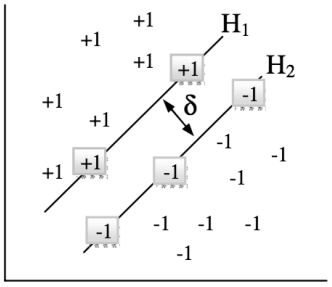

Figura 2. Hiperplano de separação para padrões lineares. Fonte: [Soares 2008]

Para conjunto de dados com padrões não-linearmente separáveis, o classificador SVM mapeia os padrões de entrada em um vetor de características com alta dimensão, para que a separação ocorra no novo espaço. Porém, a tarefa de encontrar uma função de transformação de espaço não é trivial. Para resolver o problema, a transformação ocorre com o auxílio de funções não-lineares que são denominadas de *Kernel*, que torna possível a construção de um hiperplano de separação ótimo para o espaço de características sem considerar explicitamente o espaço [Soares 2008].

Com esse espaço definido, o SVM procura o hiperplano de margem máxima (*MMH - Maximal Margin Hyperplane*), que separa os padrões no conjunto de treinamento, usando as teorias de dimensão VC (*Vapnik - Chervonenkis*) e minimização do risco estrutural [Lorena and de Carvalho 2007].

3. Metodologia

A metodologia de diagnóstico de melanoma proposta é apresentada na Figura 3.

Foi utilizado o banco de dados da organização de Colaboração Internacional de Imagens de Pele (ISIC) [Gutman et al. 2016] composto por 900 imagens dermatoscópicas, sendo 171 da classe melanoma e 729 da classe não-melanoma, e 900 máscaras binárias, criadas de forma semiautomáticas e com auxílio de especialistas, cobrindo as regiões de interesses (lesões). Utilizar a base toda demandaria muito tempo computacional; então, apenas para validar a metodologia, foi utilizada uma sub-base, selecionada aleatoriamente, 85 melanoma e 130 normais, totalizando 215 indivíduos.

Com a imagem, aplicamos a máscara fornecida pela base para delimitar a região do melanoma. Em seguida, é realizada a minimização de ruídos indesejáveis através de um filtro gaussiano. Por último, foram testados os esquemas de cores monocromático (em escala de cinza - *grayscale*), RGB e HSV. Em cada esquema foram aplicadas um dos

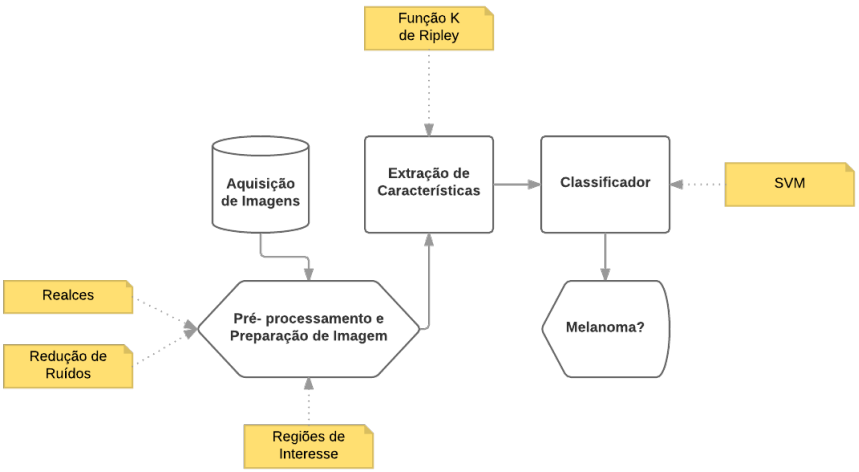

Figura 3. Metodologia proposta para o diagnóstico de melanoma.

pré-processamentos seguintes: equalização de histograma ou realce logarítmico. Desta forma, busca-se averiguar qual dos pré-processamentos em conjunto com qual esquema melhor auxilia o processo de reconhecimento proposto por este trabalho.

Com as imagens já pré-processadas, são extraídas as características usando a Função K de Ripley, em duas abordagens espaciais: uma em círculos e outra em anéis. Um pixel é considerado como um indivíduo e seu valor forma uma determinada espécie, a qual é analisada ao longo de todo o espaço cartesiano que forma a imagem e a área considerada. Foram gerados 3 círculos e 3 anéis concêntricos para cada imagem, que passa por um processo de quantização para os níveis 256, 128, 64, 32, 16 e 8. O comprimento dos raios são ajustados ao tamanho da região suspeita e as medidas de Ripley são calculadas para cada raio, em cada quantização diferente e em cada canal da imagem. Os valores finais são agrupados para formar o vetor de características cujo resumo pode ser visto na Tabela 1.

Tabela 1. Número de características geradas em cada configuração abordada.

Esquema (canais)	Aréa	Qtd de Raios	Total de Cores	Características
Grayscale (1)	Circular	3	504	1512
Grayscale (1)	Aneis	3	504	1512
RGB (3)	Circular	3	504	4536
RGB (3)	Aneis	3	504	4536
HSV (3)	Circular	3	504	4536
HSV (3)	Aneis	3	504	4536

O algoritmo de aprendizado de máquina, neste estudo, foi treinado primeiramente com 70% e depois com 80% com uma amostra de 215 da base de dados, mantendo ainda uma proporção de 1:1 em relação a casos de melanoma e não melanoma. O restante da amostra é usado para teste.

Vale ressaltar que, em cada experimento, novos conjuntos aleatórios e disjuntos

são gerados para compor o treinamento e a validação. Para avaliar o classificador em relação à sua capacidade de generalização, os critérios de acurácia, precisão, sensibilidade, especificidade e *f-score* foram analisados para cada configuração escolhida na etapa de extração de características.

4. Discussão e Resultados

Na Tabela 2, são apresentas as médias dos resultados obtidos com análise espacial através da Função K de Ripley. A primeira coluna mostra o esquema de cores utilizado, na segunda, porcentagem usada no treinamento. Já na terceira coluna, configuração, existem dois valores: o primeiro indica o tipo de área utilizada para análise (1 - área circular e 2 - em anéis), enquanto que o segundo valor indica o tipo de realce utilizado (1 - equalização de histograma e 2 - realce logarítmico). Nas demais, Ac - acurácia, P - precisão, E - Especificidade, S - Sensibilidade, Fs - *f-score* e nSV - número de vetores de suporte. Para cada uma dessas medidas são apresentados os valores da média e desvio padrão encontrados a partir de 5 experimentos de cada configuração. Pode-se observar que a melhor configuração foi alcançada pelo esquema HSV ao obter 87% de precisão e 76% de sensibilidade, usando área circular e realce logarítmico no componente V da imagem.

Tabela 2. Resultados obtidos através de análise espacial com Função K de Ripley

Esquema	T (%)	Config	Ac	P	E	S	Fs	nSV
GRAYSCALE	70%	(1,1)	0,70 ± 0,05	0,81 ± 0,03	0,88 ± 0,03	0,52 ± 0,12	0,63 ± 0,09	110,67 ± 7,36
		(1,2)	0,77 ± 0,07	0,76 ± 0,08	0,75 ± 0,09	0,79 ± 0,08	0,77 ± 0,07	109,00 ± 3,56
		(2,2)	0,78 ± 0,05	0,80 ± 0,07	0,81 ± 0,08	0,75 ± 0,05	0,77 ± 0,05	81,67 ± 7,13
		(2,1)	0,69 ± 0,05	0,76 ± 0,10	0,80 ± 0,12	0,57 ± 0,04	0,65 ± 0,03	113,67 ± 8,96
	80%	(1,1)	0,70 ± 0,01	0,75 ± 0,03	0,80 ± 0,03	0,59 ± 0,00	0,66 ± 0,01	121,00 ± 2,45
		(1,2)	0,70 ± 0,01	0,69 ± 0,02	0,67 ± 0,06	0,73 ± 0,07	0,70 ± 0,03	107,00 ± 14,90
		(2,1)	0,71 ± 0,06	0,79 ± 0,12	0,82 ± 0,13	0,59 ± 0,05	0,67 ± 0,05	136,00 ± 0,00
		(2,2)	0,75 ± 0,03	0,79 ± 0,05	0,80 ± 0,06	0,71 ± 0,00	0,74 ± 0,02	113,33 ± 9,81
RGB	70%	(1,2)	0,74 ± 0,07	0,76 ± 0,11	0,75 ± 0,15	0,73 ± 0,02	0,74 ± 0,04	95,00 ± 4,32
		(1,1)	0,70 ± 0,03	0,70 ± 0,07	0,68 ± 0,12	0,72 ± 0,09	0,70 ± 0,03	96,33 ± 1,25
		(2,1)	0,71 ± 0,07	0,71 ± 0,10	0,69 ± 0,13	0,72 ± 0,03	0,71 ± 0,05	97,33 ± 5,31
		(2,2)	0,69 ± 0,02	0,70 ± 0,00	0,72 ± 0,03	0,67 ± 0,08	0,68 ± 0,05	93,33 ± 4,99
	80%	(1,1)	0,75 ± 0,04	0,75 ± 0,02	0,75 ± 0,03	0,76 ± 0,08	0,76 ± 0,05	103,00 ± 4,32
		(1,2)	0,71 ± 0,02	0,73 ± 0,02	0,75 ± 0,06	0,67 ± 0,10	0,69 ± 0,05	102,33 ± 5,19
		(2,1)	0,72 ± 0,05	0,80 ± 0,08	0,84 ± 0,07	0,59 ± 0,08	0,67 ± 0,06	95,00 ± 8,83
		(2,2)	0,73 ± 0,06	0,74 ± 0,09	0,73 ± 0,14	0,73 ± 0,03	0,73 ± 0,03	113,00 ± 16,31
HSV	70%	(1,2)	0,71 ± 0,03	0,72 ± 0,02	0,75 ± 0,02	0,67 ± 0,08	0,69 ± 0,05	94,33 ± 0,94
		(1,1)	0,61 ± 0,05	0,62 ± 0,06	0,64 ± 0,09	0,57 ± 0,05	0,59 ± 0,04	100,33 ± 3,77
		(2,2)	0,75 ± 0,02	0,74 ± 0,05	0,72 ± 0,09	0,77 ± 0,07	0,75 ± 0,02	101,33 ± 6,13
		(2,1)	0,65 ± 0,05	0,67 ± 0,03	0,71 ± 0,04	0,60 ± 0,12	0,63 ± 0,07	112,33 ± 5,79
	80%	(1,2)	0,82 ± 0,05	0,87 ± 0,05	0,88 ± 0,05	0,76 ± 0,08	0,81 ± 0,05	99,00 ± 9,42
		(1,1)	0,75 ± 0,04	0,84 ± 0,13	0,84 ± 0,15	0,65 ± 0,08	0,72 ± 0,02	115,67 ± 6,94
		(2,2)	0,76 ± 0,00	0,80 ± 0,03	0,82 ± 0,05	0,71 ± 0,05	0,75 ± 0,01	111,67 ± 4,64
		(2,1)	0,73 ± 0,05	0,69 ± 0,04	0,65 ± 0,05	0,80 ± 0,07	0,74 ± 0,05	122,00 ± 3,56

Ao examinar os resultados, verifica-se que o esquema em tons de cinza obteve desempenho muito baixo, enquanto que o HSV demostrou alto potencial. Neste caso, a técnica analisa os canais separadamente, caracterizando os níveis de intensidade de cada canal com um determinado valor. O esquema RGB, além de possuir interdependência entre os canais, sofreu pré-processamento em cada canal, que influenciou na cor original. No HSV, apenas o canal de luminosidade foi utilizado para processamento, mantendo a originalidade da cor propriamente dita, o que levou o esquema aos resultados apresentados.

5. Conclusão

Este trabalho apresentou uma metodologia em construção para detecção de câncer de pele melanoma em imagens dermatoscópicas usando análise geoestatística com intuito de indicar padrões entre malignidade e benignidade de lesões dermatológicas a fim de serem classificadas através de máquina de vetores de suporte. A metodologia foi aplicada em uma amostra da base pública de dermatoscopias ISIC, apresentando resultados promissores, com valores de 87% e 76% para precisão e sensibilidade.

Vale ressaltar que apenas uma parte da base foi utilizada, não permitindo a completa comparação do método proposto com o estado da arte. Dado o exposto, faz-se necessário estender o desenvolvimento do trabalho com outras atividades: Aplicar a metodologia à base ISIC completa e a outras bases utilizadas nos trabalhos relacionados, explorar outros métodos de aprendizado de máquina, como redes neurais profundas e estratégias *ensemble* e implementar uma ferramenta CAD móvel para auxiliar no diagnóstico precoce da doença.

Referências

Bakheet, S. (2017). An svm framework for malignant melanoma detection based on optimized hog features. *Computation*, 5(1):4.

Braz Junior, G., da Fonseca Neto, J. V., Silva, A. C., and Paiva, A. C. (2014). *Detecção de Regiões de Massas em Mamografias usando índices de Diversidade, Geoestatística e Geometria Côncava*. PhD thesis, Universidade Federal do Maranhão.

de Oliveira Martins, L., Junior, G. B., da Silva, E. C., Silva, A. C., and de Paiva, A. C. (2007). Classification of breast tissues in mammogram images using ripley's k function and support vector machine. In *International Conference Image Analysis and Recognition*, pages 899–910, Canada. Springer.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*.

Haase, P. (1995). Spatial pattern analysis in ecology based on ripley's k-function: Introduction and methods of edge correction. *Journal of vegetation science*, 6(4):575–582.

Lorena, A. C. and de Carvalho, A. C. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67.

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S., Jafari, M., Ward, K., and Najarian, K. (2016). Melanoma detection by analysis of clinical images using convolutional neural network. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference*, pages 1373–1376, Orlando. IEEE.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212.

Soares, H. B. (2008). *Análise e classificação de imagens de lesões da pele por atributos de cor, forma e textura utilizando máquina de vetor de suporte*. PhD thesis, Universidade Federal do Rio Grande do Norte.