MODELOS PARA DIFERENCIAÇÃO DE TECIDOS CÓLICOS EM IMAGENS DE COLOSCOPIA

Jefferson Tales Oliva[1], Carlos Andrés Ferrero[1]

[1]Centro de Engenharias e Ciências Exatas (CECE), Universidade Estadual do Oeste do Paraná – UNIOESTE, Foz do Iguaçu (PR), Brasil

Resumo: Esse trabalho tem como objetivo construir modelos de classificação para diferenciar fragmentos de imagens representativas de tecidos cólicos utilizando técnicas de Análise de Imagens e Aprendizado de Máquina. O método proposto foi aplicado em um conjunto de 67 modelos de imagem provenientes de exames de coloscopia, contendo pólipos. Nesse sentido, os fragmentos de imagem foram representados mediante características extraídas da Matriz da Diferença dos Tons de Cinza da Vizinhança e a classificação de exemplos foi realizada utilizando o algoritmo J48 e a técnica Nearest Neighbor. Os resultados evidenciaram que o modelo construído com J48 apresentou melhor desempenho em relação à técnica de classificação Nearest Neighbor, para as quatro medidas analisadas. De acordo com os especialistas do domínio, os modelos construídos se mostraram promissores para o estudo de padrões de imagem de tecido cólico, no entanto, mais estudos necessitam ser desenvolvidos.

Palavras-chave: Neoplasias Colorretais, Diagnóstico por Imagem, Inteligência Artificial.

Abstract*: To build classification models to differentiate colon tissues in representative coloscopy images by Image Analysis and Machine Learning techniques. The study was conducted through four steps: select tissues area into coloscopy images; extract features, build computational models and predictive analysis. The method was applied to a set of 67 coloscopy images models, containing polyps. All images were represented by the image features extracted from a respective Neighborhood Gray-Tone Difference Matrix, and then, classification of samples was performed using the J48 algorithm and Nearest Neighbor technique. Results show that the model built with J48 has a better performance compared to the Nearest Neighbor classification technique, for the four measures evaluated. Built models were promising to study of colon tissue image patterns, however, further studies are needed.*

Keywords*: Neoplasms, Diagnostic Imaging, Artificial Intelligence*.

Introdução

Em hospitais e clínicas médicas, grande quantidade de dados provenientes de exames médicos são armazenados frequentemente em formatos multimídia (vídeo, texto e áudio)[1]. Especificamente, as imagens médicas são armazenadas no intuito de manter os registros de exames ao longo do tempo para auxiliar os especialistas da área médica no diagnóstico de enfermidades em pacientes.

De acordo com Instituto Nacional do Câncer (INCA), no Brasil, o câncer colorretal constitui a terceira neoplasia maligna de maior incidência em homens e a segunda em mulheres, sendo que a sobrevida média global é de cinco anos após sua detecção precoce[2]. Nesse sentido, o exame de coloscopia se apresenta como uma ferramenta indispensável para o diagnóstico de enfermidades do intestino grosso, como pólipos e úlceras, que podem resultar no surgimento dessa doença[3].

A utilização de técnicas computacionais para analisar as imagens provenientes desse exame pode auxiliar especialistas da área médica no aprimoramento do diagnóstico e no melhor entendimento dessas enfermidades. Nesse sentido, esse trabalho constitui parte do projeto de característica multidisciplinar denominado Análise de Imagens Médicas, o qual está sendo desenvolvido no Laboratório de Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE / Foz do Iguaçu) em parceria com o Serviço de

Coloproctologia da Faculdade de Ciências Médicas (FCM) da Universidade Estadual de Campinas (UNICAMP). Esse projeto tem como propósito desenvolver ferramentas computacionais para a análise inteligente de dados médicos no formato de imagem.

Sendo assim, o objetivo desse trabalho consiste em utilizar técnicas de Análise de Imagens (AI) e de Aprendizado de Máquina (AM) para construir modelos de classificação que permitam diferenciar fragmentos de imagem de tecido cólico de pólipo e sem pólipo.

Métodos

Para realização desse trabalho foi desenvolvido um método direcionado ao estudo de padrões em fragmentos de imagem, que é constituído de quatro etapas: (1) seleção de fragmentos, (2) extração de características, (3) construção de modelos e (4) análise de resultados.

Etapa 1: seleção de fragmentos

Sistemas de gerenciamento de imagens vêm se tornando cada vez mais comuns em hospitais e clínicas médicas, no intuito de prover suporte ao armazenamento de exames nesse formato. No contexto de exames de coloscopia, as imagens registram anormalidades relacionadas ao intestino grosso, como do tipo polipoide, diverticular, vascular, entre outras[3]. Nessa etapa, a partir de um conjunto de *n* imagens de tecido cólico com pólipo são selecionados dois fragmentos: um de tecido de pólipo e outro de tecido normal. Desse modo, é constituído o Conjunto de Imagens, denotado por $CI = \{Im_1, Im_2, ..., Im_{2n}\}$.

Etapa 2: extração de características

Nessa etapa, a *i-ésima* imagem do *CI* é submetida a um processo de transformação para níveis de cinza, originado o exemplo Im'_i. O conjunto de imagens normalizadas constitui o conjunto $CI' = \{Im'_1, Im'_2, ..., Im'_{2n}\}$. Para cada imagem de *CI'* é realizada a extração de características. Essas características são frequentemente relacionadas à cor, à textura ou à forma da imagem. Um dos métodos utilizados para representação de características de textura consiste na construção da Matriz da Diferença dos Tons de Cinza da Vizinhança (MDTCV)[4], a qual é representada pela diferença entre tons de cinza de cada pixel da imagem e a média dos tons de cinza dos seus vizinhos. O *j-ésimo* componente da MDTCV resume a diferença entre os *pixels* com tom de cinza *j* e o tom de cinza médio de seus vizinhos, considerando todos os pixels de uma imagem, exceto os *pixels* situados na borda em um raio de distância *d*. A partir das MDTCV, podem ser extraídas cinco características[4], que são representadas por valores numéricos, tais como: aspereza, fineza, complexidade, força de textura e contraste. Na Figura 1 é apresentado um exemplo de imagem e a MDTCV correspondente, onde o valor de *j* corresponde ao nível de cinza e *S(j)* corresponde à soma da diferença entre os pixels com valor *j* e o valor dos *pixels* vizinhos, considerando uma distância *d=1*. Nessa figura, ao lado direito, se destacam em formato negrito todos os *pixels* que são considerados para avaliar a respectiva vizinhança e constituir a MDTCV, ao lado esquerdo.

Desse modo, cada imagem do conjunto *CI'* é representada por um conjunto de características, que consistem em valores calculados a partir da MDTCV construída para cada imagem.

Imagem

3	2	0	1	0
1	**2**	**1**	**3**	0
3	**1**	**0**	**2**	3
1	**2**	**3**	**0**	3
0	0	0	0	1

Matriz da Diferença dos Tons de Cinza da Vizinhança

j	*S(j)*
0	3,25
1	1,00
2	2,00
3	4,50

Figura 1: Exemplo de imagem representada como uma matriz de *pixels 5X5* e a sua respectiva MDTCV para distância igual a *1*[5].

Etapa 3: construção de modelos

Após a extração de características, é realizada a construção de modelos para classificação de exemplos por meio da aplicação de técnicas de AM. Dentre as diferentes técnicas para esse fim se destacam: (1) árvores de decisão e (2) vizinhos mais próximos. O método (1) consiste na construção de uma estrutura de dados organizada hierarquicamente, baseada nos conceitos de divisão e conquista. Posteriormente, essa estrutura de dados é percorrida desde a raiz até as folhas para classificação de novos exemplos[6]. O método (2) consiste na classificação de novos exemplos por meio do cálculo da similaridade desse exemplo com exemplos de treinamento, previamente classificados por especialistas[7]. Desse modo, a partir dos exemplos mais similares é definida a classe do novo exemplo por meio de uma função que determina a classe, considerando as classes dos exemplos mais similares. É importante ressaltar que nessa abordagem não é construído um modelo estrutural específico, considerando o conjunto de exemplos de treinamento o próprio modelo.

Etapa 4: análise de resultados

Na quarta etapa, os modelos são avaliados quanto à qualidade preditiva, considerando a eficiência para classificação de novos exemplos. A opinião dos especialistas do domínio também deve ser considerada, com o intuito de definir modelos, a partir dos quais, seja possível futuramente a construção de sistemas especialistas.

Um dos métodos de análise de resultados consiste na construção de Tabelas de Contingência (TC), que são utilizadas para avaliar o relacionamento entre duas ou mais variáveis nominais, isto é, se pertence ou não pertence a uma classe. Por exemplo, em se tratando da classificação de fragmentos de imagem de tecido cólico, estes podem pertencer à classe de tecido de anormalidade ou sem anormalidade. A partir da TC podem ser extraídas quatro medidas de precisão[8], as quais são definidas a seguir:

- **Valor Preditivo Positivo (VPP)**: define a percentagem de fragmentos de imagem de tecido cólico com anormalidade em relação ao total de exemplos classificados como anormais;
- **Valor Preditivo Negativo (VPN)**: define a percentagem de fragmentos de imagem de tecido cólico sem anormalidade em relação ao total de exemplos classificados como sem anormalidade;
- **Sensibilidade**: define a probabilidade de um fragmento de imagem de tecido cólico com anormalidade ser classificado como anormal;
- **Especificidade**: define a probabilidade de fragmentos de imagem de tecido cólico sem anormalidade ser classificado como sem anormalidade.

As Etapas (1) e (2) do método apresentado são realizadas com auxílio da ferramenta computacional *Medical Imaging Analysis System* (MIAS)[9], desenvolvida no Laboratório de

Bioinformática (LABI) da Universidade Estadual do Oeste do Paraná (UNIOESTE / Foz do Iguaçu). As etapas (3) e (4) são realizadas utilizando a ferramenta WEKA[10] para construção de modelos e o software GraphPad Instat® para a análise estatística. A ferramenta WEKA oferece diversos algoritmos de AM, tais como: o J48 para construção de árvores de decisão, que é uma implementação do algoritmo C4.5 proposto por Quinlan[11]; e o NN (*Nearest-Neighbor*) para classificação pelos vizinhos mais próximos, que armazena todas as instâncias de treinamento e calcula a similaridade de novos exemplos por meio da distância Euclidiana[10].

Resultados

O método apresentado foi aplicado a um conjunto de 67 imagens representativas de exames de coloscopia provenientes do Serviço de Coloproctologia da Faculdade de Ciências Médicas da UNICAMP. Esses modelos de imagem representam tecidos cólicos de pólipo Tipo Ip (Protruso Pediculado, de acordo com a classificação de morfologia de pólipos da Sociedade Japonesa de Pesquisa do Câncer Colorretal)[3].

Na Etapa 1, para cada imagem, foram selecionados manualmente em conjunto com os especialistas do domínio, dois fragmentos de imagem, sendo um que representa tecido de pólipo e outro que representa tecido sem pólipo, constituindo o conjunto *CI*, contendo 134 fragmentos de imagem.

Na Etapa 2, com auxílio do aplicativo MIAS, cada fragmento de imagem foi transformado do formato RGB (*Red-Green-Blue*) para o formato de níveis de cinza, considerando 64 níveis. Logo, para cada fragmento de imagem em tons de cinza foi construída a MDTCV, considerando *d=1* e em seguida, extraídas cinco características propostas em[4]: aspereza, fineza, complexidade, força de textura e contraste.

Na Etapa 3, mediante o aplicativo WEKA, foram construídos os modelos de classificação por meio dos algoritmos J48 para construção de árvores de decisão e NN para classificação por meio da técnica de vizinhos mais próximos.

Após, na Etapa 4, os modelos foram avaliados com base na precisão preditiva, mediante a análise dos resultados pelo uso de TC. Desse modo, nas Tabelas 1 e 2 são apresentadas as TC construídas para cada um dos modelos.

Tabela 1 – TC para o modelo construído por meio do algoritmo J48.

Pólipo	**Classificação**		**Total**
	Presente	**Ausente**	
Presente	55	17	72
Ausente	12	50	62
Totais	67	67	134

Tabela 2 – TC para o modelo construído por meio do algoritmo NN.

Pólipo	**Classificação**		**Total**
	Presente	**Ausente**	
Presente	44	21	65
Ausente	23	46	69
Totais	67	67	134

Sendo assim, a partir dos resultados de cada TC foram calculadas as medidas de precisão, que são apresentadas na Tabela 3.

Tabela 3 – Medidas Calculadas a partir das TC dos modelos de classificação.

Algoritmo	VPP (%)	VPN (%)	Sen. (%)	Esp. (%)
J48	82,09	74,63	76,39	80,65
1NN	65,67	68,66	67,69	66,67

Discussão

De acordo com os dados apresentados nas tabelas 1 e 2, é possível observar que o modelo construído através do algoritmo J48 obteve melhor desempenho para classificar imagens de tecidos cólicos de pólipo em relação ao algoritmo NN, sendo 55 fragmentos de imagem presença de anormalidade classificados corretamente e 50 sem presença de anormalidade.

Por meio da Tabela 3, é constatado que o modelo construído por meio do algoritmo J48 obteve os maiores valores para todas as medidas avaliadas, como VPP, VPN, Sensibilidade e Especificidade, sendo 82,09%, 74,63%, 76,39% e 80,65%, respectivamente. Assim, esse modelo apresentou melhor desempenho e maior probabilidade de classificação correta de fragmentos imagem representativos de tecidos cólicos, em relação ao algoritmo NN.

Em trabalho anterior[12], os modelos de classificação construídos por meio dos algoritmos J48 e NN foram avaliados por meio do cálculo de média e desvio-padrão resultante do processo de validação cruzada[1]. Para verificar a existência de diferença estatisticamente significativa entre esses dois modelos, foi aplicado o teste estatístico *t-student* para dados emparelhados, considerando nível de significância de 95%, por meio do qual, não foi possível constatar diferença estatisticamente significativa entre ambos os modelos. No obstante, nesse trabalho, mediante a avaliação por TC, foi evidenciado que o modelo construído com o algoritmo J48 apresentou melhor desempenho para classificação de fragmentos de imagem de tecidos cólicos em relação à classificação realizada pelo algoritmo NN nas quatro medidas preditivas avaliadas.

Conclusão

Nesse trabalho foi apresentado um método de construção e avaliação de modelos de classificação para diferenciação de fragmentos de imagem de tecido cólico de anormalidade e sem anormalidade, a partir da representação desses fragmentos por meio de atributos de textura. Para essa representação foi utilizada a Matriz da Diferença dos Tons de Cinza da Vizinhança e foram extraídas cinco características de suas respectivas matrizes. Posteriormente, foram utilizados os algoritmos de Aprendizado de Máquina denominados J48 e *Nearest-Neighbor,* para a construção de modelos e classificação. Por fim, foram usados dois métodos para avaliação de modelos, validação cruzada e tabela de contingência.

Embora o modelo construído pelo algoritmo J48 não tenha apresentado diferença estatisticamente significativa em relação ao algoritmo *Nearest-Neighbor*, na análise por meio da Tabela de Contingência foi possível constatar que o modelo construído pelo algoritmo J48 obteve maiores valores preditivos para Valor Preditivo Negativo, Valor Preditivo Positivo, Sensibilidade e Especificidade em relação ao modelo construído mediante ao algoritmo *Nearest-Neighbor*. Nesse contexto, de acordo com a opinião dos especialistas, os modelos se apresentam promissores para o estudo de imagens de tecidos cólicos, no entanto, mais estudos necessitam serem desenvolvidos no intuito de evidenciar técnicas de representação e algoritmos de construção de modelos que permitam identificar de modo mais completo os padrões de tecidos cólicos normais e de anormalidade.

Desse modo, trabalhos futuros incluem representação de imagens por meio de outros atributos de textura; construção de modelos por meio de outros algoritmos de Aprendizado de Máquina; e comparação de modelos construídos por meio de atributos de texturas diferentes.

Agradecimentos (opcional)

À Fundação Araucária e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CNPq) pelo apoio por meio da linha de apoio a projetos de pesquisa e de concessão de bolsa de pós-graduação na categoria mestrado, respectivamente.

Referências

[1] Karkanis S, Galoussi K, Maroulis D. Classification of endoscopic images based on texture spectrum. Proceedings of Workshop on Machine Learning in Medical Applications; 1999 Jul 63-69; Chania, Greece. Pennsylvania: CitSeerX; 1999.
[2] Instituto Nacional do Câncer - INCA. Estimativa 2012: Incidência de Câncer no Brasil. 2011 Nov [citado 2012 mai 31]. Disponível em: http://www,inca.gov.br/estimativa/2012/estimativa20122111.pdf
[3] Quilici F. Coloscopia. São Paulo: Editora Lemos; 2000.
[4] Amadasum M, King R. Textural Features Corresponding to Textural Properties. IEEE Transactions on Systems, Man, and Cybernetics. 1989; 19(5): 1264-10.
[5] Pedrini H, Schwartz WR. Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações. São Paulo: Thomson; 2008.
[6] Rezende, S. Sistemas Inteligentes: Fundamentos e Aplicações. Barueri: Editora Manole; 2003.
[7] Alpaydin, E.Introduction to Machine Learning. Cambridge: MIT Press; 2004.
[8] Freedman D, Pisani R, Purvers R. Statistics. New York: Norton; 1998.
[9] Ferrero CA, Lee HD, Spolaôr N, Coy CRS, Fagundes JJ, Machado RB, et al. Estudo comparativo de modelos computacionais gerados sobre representações de imagens de coloscopia: tecido de mucosa normal VS tecido de mucosa de pólipo cólico. Revista Brasileira de Coloproctologia. 2009; 29(1): 23-7.
[10] Witten I, Frank E. Machine Learning: Practical Machine Learning Tools and Techniques. São Francisco: Morgan Kaufmann; 2005.
[11] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco: 1993.
[12] Oliva JT, Ferrero CA, Coy CSR, Fagundes JJ, Machado, RB, Lee, HD, et al. Construção de modelos de classificação de fragmentos de tecido cólico de pólipo e sem pólipo por da representação de imagens por atributos de textura. In: Anais do XIX Encontro Anual de Iniciação Científica; 2010; out. 28 – 30; Guarapuava. Paraná. [Internet] [citado 2012 mai 31]. Disponível em: http://anais.unicentro.br/xixeaic/pdf/834.pdf

Contato

J. T. Oliva — jeffersontalesoliva@gmail.com
C. A. Ferrero — anfer86@gmail.com

Laboratório de Bioinformática — LABI, Universidade Estadual do Oeste do Paraná — UNIOESTE, Av. Tancredo Neves, 6731, CEP 85866-900, Foz do Iguaçu — PR.