

Avaliação de Algoritmos de Classificação de Padrões na Detecção de Câncer do Colo do Útero

Eduardo F. Montezuma, Levi P. S. A. Alencar, Guilherme A. Barreto

Universidade Federal do Ceará (UFC), Departamento de Engenharia de Teleinformática

Bloco 725, Centro de Tecnologia, Campus do Pici, Fortaleza, Ceará

e-mails: edumachine@hotmail.com, levipioalencar@gmail.com, gbarreto@ufc.br

Abstract - *The main goal of this study is to empirically evaluate linear and nonlinear pattern classifiers in the detection of cervical cancer. For this purpose, this paper reports a comprehensive performance comparison study involving 4 different classification methods, namely: linear least squares classifier, extreme learning machine, support vector machine and quadratic Gaussian classifiers. All the classifiers are evaluated on a benchmarking database generated from pap-smear image analysis. The results of the experiments indicate that the linear classifiers perform as good as powerful nonlinear ones, with the advantage of demanding much lower computational resources.*

Palavras-chave: *Exame Papanicolau, câncer cervical, classificação de padrões, SVM, ELM.*

Introdução

O câncer de colo de útero, conhecido também como câncer cervical, é um dos tipos de carcinoma que mais atingem a população feminina, perdendo em nível de incidência apenas para os cânceres de mama e colorretal. Um diagnóstico rápido e eficaz é de fundamental importância para a efetividade de tratamentos em casos de diagnóstico positivo. O termo *Pap-Smear* se refere ao esfregaço de Papanicolau, que é um exame de triagem para o câncer de colo uterino.

As amostras de células são esfregadas em uma lâmina de vidro e coloridas, tornando mais fácil de examinar as células sob um microscópio, procurando anormalidades que indiquem um estágio pré-cancerígeno. Este exame não deve ser encarado como um método de diagnóstico, mas sim como forma de selecionar pacientes com risco de apresentarem lesões pré-neoplásicas e invasoras, devendo ser submetidas a processos de acompanhamento estritos e a outros exames complementares.

O propósito do Teste Papanicolau é diagnosticar células pré-cancerígenas antes que elas progridam para um *carcinoma invasivo* [1]. De fato, é necessário um citotécnico experiente para diferenciar

entre os diferentes tipos de células. Cada lâmina contém cerca de 300.000 células de modo que a classificação é um trabalho demorado, portanto sistemas semi-automáticos têm sido desenvolvidos para ajudá-los [2].

Em anos recentes, tem se observado uma tendência de se utilizar técnicas avançadas de processamento de imagens e aprendizado de máquinas visando aumentar a efetividade do diagnóstico médico de câncer cervical [3, 4]. Este tipo de sistema de auxílio ao diagnóstico pode ser útil em treinamento de pessoal especializado e, principalmente, suprir carências de citotécnicos em regiões afastadas dos grandes centros urbanos.

No presente artigo é apresentado um estudo comparativo do desempenho de métodos estatísticos, matemáticos e neurais de classificação de padrões na identificação de anormalidade nas células anormais do colo do útero obtidas via teste Papanicolau. São avaliados 4 algoritmos, a saber: classificador linear de mínimos quadrados, máquina de aprendizado extremo, máquina de vetores de suporte e classificador gaussiano quadrático. A literatura da área é carente em tais estudos, sendo esta a principal motivação para o desenvolvimento deste trabalho.

Materiais e Métodos

O conteúdo textual desta seção é adaptado da referência [2], que é o trabalho responsável pela disponibilização do conjunto de dados usado neste artigo. Preferencialmente, as amostras são tomadas de várias áreas do colo do útero (Fig. 1). Elas, na maioria dos casos, contém células do epitélio colunar e escamoso. O epitélio colunar encontra-se na parte superior do colo do útero, enquanto o escamoso, se encontra na parte inferior. Entre estes encontra-se o epitélio metaplásico, também chamado de zona de transformação, ou junção escamo-colunar.

No epitélio escamoso, existem 4 camadas de células. As células formam a camada *basal*, e enquanto maturam, elas se movem para cima

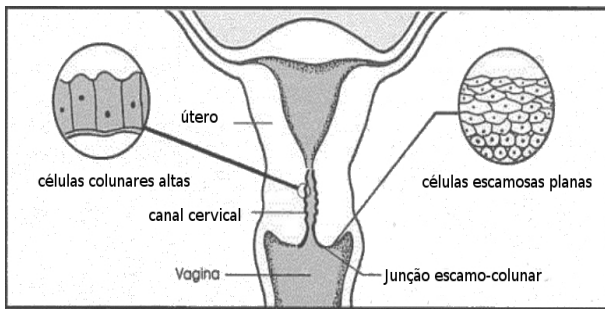


Figura 1 – Localização de células colunares e escamosas. Células escamosas desenvolvem-se de modo ascendente a partir da camada basal para as camadas parabasal, intermediária e superficial. Displasia desenvolve-se na junção escamo-colunar. Adaptado de [2].

pela camada *parabasal*, a camada *intermediária*, e finalmente a camada *superficial*. Aquelas que encontram-se na camada *basal* dividem-se e fornecem células às camadas acima. Enquanto elas movem-se através das camadas, as células mudam de formato, cor e outras características. Ao chegarem na camada *superficial*, elas são rejeitadas e substituídas por células vindas de baixo. A camada *basal* tem células pequenas e redondas, com um núcleo relativamente grande, e um citoplasma pequeno. Ao amadurecer, o núcleo se torna menor, e o citoplasma torna-se maior. A forma da célula se torna menos arredondado conforme ela matura.

O epitélio colunar possui somente uma camada de células, contendo dois tipos: colunares e reserva. As células reserva se dividem em novas células reserva, e em novas células colunares. Nas células colunares normais, o núcleo é localizado na parte inferior do citoplasma. Quanto vistas pelo topo, o núcleo parece maior enquanto quando vistas pelo lado, o citoplasma parece maior. O epitélio metaplásico consiste de células reserva do epitélio colunar. Quando as células amadurecem por completo, elas parecem células obtidas no epitélio escamoso.

Nas células displásticas, a informação genética foi alterada de alguma forma, e a célula não divide-se mais do que deveria. Esta é uma célula pré-cancerígena. Dependendo do tipo de célula que se divide incorretamente, podemos adquirir diagnósticos como *displasia* ou *carcinoma in situ*. O termo “plasia” significa crescimento, e “displasia”, crescimento desordenado. As células displásticas dividem-se em três categorias: levemente, moderadamente ou severamente displásticas. A classificação é determinada à partir da probabilidade de as células tornarem-se células cancerígenas

malignas. Uma grande quantidade de células suavemente displásticas irão desaparecer sem se tornar malignas, enquanto as severas, provavelmente tornar-se-ão células malignas. Células displásticas escamosas geralmente tem um núcleo maior e mais escuro, e tendem a formar aglomerados. Em displasias severas, os núcleos são grandes, com grânulos negros e usualmente deformados. O citoplasma é pequeno e negro comparado ao núcleo.

Carcinoma cervical é o tipo mais comum de malignidade no sistema genital feminino. A carcinogênese é um processo longo, que começa num epitélio normal, que torna-se displástico, evolui para *carcinoma in situ*, e então para câncer. O longo intervalo entre os estágios permite a possibilidade de um diagnóstico precoce, com elevada possibilidade de cura completa.

Banco de Dados: O banco de dados *Pap-smear* é o último de duas versões construídas pelo Hospital Universitário de Herlev, afiliado à Universidade de Copenhague, Dinamarca. As imagens foram preparadas e analisadas por citotécnicos especializados e médicos usando uma câmera digital e um microscópio. Um pacote comercial de software foi usado para segmentar as imagens. Cada célula foi examinada por dois citotécnicos, e amostras difíceis também por um médico. Em caso de divergência, a amostra foi descartada. O banco de dados, então, possui diagnósticos que são tão precisos quanto possível, dadas as limitações práticas e econômicas do hospital.

As células foram selecionadas, não para coletar uma distribuição natural, mas para criar uma boa coleção das classes importantes. Existem pelo menos o dobro de células anormais, em relação às células normais; a composição precisa de classes é mostrada na Tabela 1. Neste artigo, o problema de classificação em 7 classes foi convertido em um problema de classificação binária (2 classes): classe ω_1 (normal com 242 instâncias) e classe ω_2 (anormal, com 675 instâncias).

O banco de dados está disponível para uso público a partir do site do MDE-LAB¹ (*Management and Decision Engineering Laboratory*). Cada amostra é descrita por 20 atributos extraídos de imagens de cada uma das células. Os atributos foram extraídos das imagens utilizando programas escritos em Matlab por Erik Martin [5]. Eles foram extraídos de uma combinação de imagens segmentadas e não-segmentadas. Os dados foram analisados extensivamente em dois projetos

¹<http://mde-lab.aegean.gr/downloads>

Classe	Categoria	Tipo de célula	Contagem de células	Subtotais
1	Normal	epitélio escamoso superficial	74	
2	Normal	epitélio escamoso intermediário	70	
3	Normal	epitélio colunar	98	242
4	Anormal	displasia leve (NIC I)	182	
5	Anormal	displasia moderada (NIC II)	146	
6	Anormal	displasia severa (NIC III)	197	
7	Anormal	célula escamosa de carcinoma <i>in situ</i> intermediário	150	675

Table 1 – Distribuição das 917 instâncias no banco de dados. Classes 1 a 3 contêm células normais, enquanto as classes 4 a 7 contêm células consideradas anormais.

de Mestrado [5, 6]. Mais detalhes sobre os atributos podem ser obtidos da leitura da referência [2].

O banco de dados, para todos os classificadores, foi dividido aleatoriamente em 80 % para treinamento e 20 % para teste. E cada rodada treino-teste foi repetida 50 vezes. Além disso, todas as amostras foram normalizadas para uma variância unitária e média nula, ou seja, os atributos de cada amostra são condicionados ao intervalo $[-1, 1]$. Os seguintes classificadores foram avaliados no presente artigo: (i) classificador linear de mínimos quadrados (MQ) [7], (ii) máquina de aprendizado extremo (ELM, do inglês *extreme learning machine*) [8], (iii) máquina de vetores de suporte (SVM, do inglês *support vector machine*) [9] e (iv) classificador quadrático gaussiano (CQG) [7].

Para todos os classificadores avaliados neste trabalho, vetores de atributos que descrevem células cancerígenas são rotulados como +1 (classe ω_1), enquanto instâncias normais recebem o rótulo -1 (classe ω_2). A plataforma computacional utilizada possui as seguintes características: Notebook Dell (Vostro 3050), com 6 GB de RAM, HD de 500 GB, processador i5 2.30GHz, rodando sistema operacional Windows 10. Os classificadores foram implementados utilizando-se as linguagens de programação *GNU Octave*² e *Julia Language*³, devido à vasta quantidade de ferramentas estatísticas e de otimização que estas possuem e também ao fato de serem linguagens livres.

Resultados e Discussão

Os resultados obtidos para 50 rodadas encontram-se na Tabela 2. Analisando-se esta tabela, é perceptível que os 3 classificadores não lineares (SVM, ELM e CQG) apresentaram variações maiores, em relação ao classificador MQ.

Ainda, pela boa performance do classificador MQ em relação aos demais modelos, pode-se afirmar que o conjunto de dados pode apresentar padrões linearmente separáveis. Vale ainda citar

que o classificador com maior acurácia foi o SVM. Pode-se atribuir isto ao seu poder de classificação. Vale ressaltar porém que, de todos os classificadores, ele foi o que mais apresentou disparidade entre o valor de acurácia mínimo e máximo. Isso pode ter sido causado devido, sobretudo, ao grande número de graus de liberdade que este classificador tem, tendo portanto grande dispersão em torno da acurácia média.

As matrizes de confusão obtidas para o melhor (*best*) e pior (*worst*) caso, respectivamente, para cada classificador avaliado são apresentadas a seguir.

1. Classificador Quadrático Gaussiano:

$$MC_{best} = \begin{bmatrix} 39 & 4 \\ 2 & 139 \end{bmatrix}, MC_{worst} = \begin{bmatrix} 34 & 17 \\ 3 & 130 \end{bmatrix}$$

2. Rede ELM:

$$MC_{best} = \begin{bmatrix} 39 & 6 \\ 1 & 138 \end{bmatrix}, MC_{worst} = \begin{bmatrix} 43 & 22 \\ 3 & 116 \end{bmatrix}$$

3. Classificador MQ:

$$MC_{best} = \begin{bmatrix} 47 & 6 \\ 1 & 130 \end{bmatrix}, MC_{worst} = \begin{bmatrix} 37 & 11 \\ 8 & 128 \end{bmatrix}$$

4. Classificador SVM:

$$MC_{best} = \begin{bmatrix} 48 & 4 \\ 0 & 132 \end{bmatrix}, MC_{worst} = \begin{bmatrix} 34 & 20 \\ 4 & 126 \end{bmatrix}$$

Para finalizar o estudo comparativo dos desempenhos dos 4 classificadores, as curvas ROC dos classificadores MQ, ELM e SVM são levantadas. Uma curva ROC é um gráfico traçado à partir da relação entre a *sensibilidade* (SENS) e *1-especificidade* (SPEC) de um classificador, variando-se o limiar (t) utilizado na classificação [10]. Através da análise das curvas ROC, é possível concluir que o classificador que demonstrou o melhor desempenho, com relação aos conceitos de sensibilidade e especificidade, foi o classificador SVM, enquanto os desempenhos dos classificadores MQ e ELM foram bem similares.

²<https://www.gnu.org/software/octave/>

³<https://julialang.org/>

Classificador/ Acurácia	Média	Mínima	Máxima	Mediana	Desvio Padrão
MQ	93.50	90.21	96.19	94.56	1.69
CQG	93.89	89.13	96.739	94.56	1.66
ELM	91.47	86.41	96.19	91.30	2.29
SVM	93.32	86.95	97.82	94.56	2.61

Table 2 – Resultados de acurácia obtidos para os classificadores avaliados.

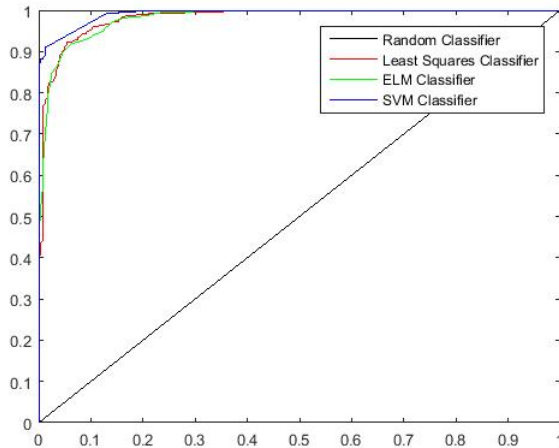


Figura 2 – Curvas ROC obtidas para os 4 classificadores avaliados.

Conclusões

Neste artigo foram reportados os resultados de um amplo estudo comparativo de desempenho envolvendo classificadores de padrões lineares e não-lineares na detecção de câncer do colo uterino. Para isso, foi usado um banco de dados de *benchmarking* construído a partir de análise de amostras de células oriundas do exame Papanicolau (*pap-smear*) de pacientes de um hospital norueguês.

Os resultados obtidos foram bem interessantes, uma vez que ficou evidenciado nos experimentos computacionais que o classificador linear de mínimos quadrados (MQ) alcançou desempenho compatível com classificadores não-lineares de elevado poder computacional, tais como SVM e rede ELM. De um modo geral, o melhor classificador foi o SVM para todas as métricas de desempenho (acurácia, matriz de confusão e curva ROC), mas este foi seguido de perto pelo classificador MQ. Os resultados obtidos foram todos melhores que os reportados em [2] para o mesmo banco de dados.

Agradecimentos

Agradecemos ao CNPq pelo suporte financeiro via bolsa PQ (309451/2015-9) e duas bolsas PIBIC (118642/2016-2 e 118635/2016-6).

Referências

- [1] A. Meisels and C. Morin, *Cytopathology of the Uterus*. ASCP Press, 1997.
- [2] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, *Pap-smear Benchmark Data For Pattern Classification*, pp. 1–9. 2005.
- [3] E. J. Mariarputham and A. Stephen, “Nominated texture based cervical cancer classification,” *Computational and Mathematical Methods in Medicine*, vol. ID 586928, pp. 1–10, 2015.
- [4] T. Xu, C. Xin, L. R. Long, S. Antani, Z. Xue, E. Kim, and X. Huang, “A new image data set and benchmark for cervical dysplasia classification evaluation.,” in *Machine Learning in Medical Imaging - MICCAI’2015* (L. Zhou, L. Wang, Q. Wang, and Y. Shi, eds.), vol. 9352 of *Lecture Notes in Computer Science*, pp. 26–35, Springer, 2015.
- [5] E. Martin, “Pap-smear classification,” Master’s thesis, Technical University of Denmark - DTU, 2003.
- [6] J. Norup, “Classification of pap-smear data by transductive neuro-fuzzy methods,” Master’s thesis, Technical University of Denmark - DTU, 2005.
- [7] A. Webb, *Statistical pattern recognition*. Wiley, 2nd ed., 2002.
- [8] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, pp. 489–501, 2006.
- [9] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [10] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.