Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Seleção de Potenciais Biomarcadores para Previsão da Recidiva do Câncer de Próstata Utilizando Modelos de Classificação

Maria E. Antunes[1], Paulo F. A. Mancera[2]
UNESP, Botucatu, SP
Thaise G. Araújo[3], Eliana Pantaleão[4], Marta H. Oliveira[5]
UFU, Patos de Minas, MG

O câncer de próstata (CaP) é o segundo mais comum entre homens no Brasil, com estimativa 71.730 novos casos para cada ano do triênio 2023-2025 [3]. Entre os principais desafios do combate à doença está o diagnóstico precoce, que é feito através do exame de toque retal e da determinação do nível de PSA (**Prostate-Specific Antigen**) no sangue. O prognóstico da doença é estabelecido através do escore de Gleason e do estadiamento TNM, que buscam descrever e avaliar o grau histológico, extensão e disseminação do tumor. Após o tratamento, o acompanhamento do paciente se dá pelas dosagens de PSA em amostras de sangue a cada trimestre ou semestre. A recidiva da doença é um problema a ser enfrentado que afeta de 20% a 30% dos pacientes, e as informações tradicionais de diagnóstico mostram-se insuficientes para prever essa recidiva de forma precoce, pois nem sempre um aumento no nível de PSA é indicativo de malignidade [2, 7].

Algoritmos de aprendizado de máquina se destacam como uma ferramenta altamente promissora para identificar padrões e resolver uma variedade de problemas, especialmente em cenários de incerteza. Eles têm a capacidade de aprender com o ambiente ao seu redor, buscando otimizar o desempenho na solução de problemas através da construção de modelos que representem conjuntos de dados relevantes [5]. Algoritmos de aprendizado supervisionado, como os de classificação, mostram-se particularmente eficazes na identificação de padrões em dados biológicos, o que pode levar a melhorias significativas na previsão de recidivas de doenças e no desenvolvimento de estratégias de tratamento mais eficazes [1, 6].

O objetivo deste trabalho foi avaliar o desempenho de potenciais biomarcadores na classificação de pacientes com CaP que apresentaram recidiva da doença, utilizando modelos de classificação de aprendizado supervisionado, como árvores de decisão e máquinas de vetores de suporte (SVM). Para isso, utilizamos um conjunto de dados do The Cancer Genome Atlas (TCGA) contendo informações de 419 pacientes com CaP. Este conjunto inclui uma ampla gama de dados, como níveis pré-cirúrgicos de PSA, escore de Gleason, estadiamento da doença e a expressão gênica de sete biomarcadores: KLK3|354, AR|367, GSTM3|294, NETO2|81831, HPN|3249, PRUNE2|158471 e FOLH1|2346 (PSMA). Durante um período de aproximadamente 7 anos, 85 dos 419 pacientes apresentaram recidiva. Portanto, o objetivo da construção dos modelos foi investigar se algum dos biomarcadores mencionados (ou uma combinação deles) poderia melhorar a assertividade na previsão de recidiva da doença, considerando que PSA, escore de Gleason e estadiamento já são variáveis utilizadas na prática clínica para diagnóstico.

[1]maria.antunes@unesp.br
[2]paulo.mancera@unesp.br
[3]tgaraujo@ufu.br
[4]epantaleao@ufu.br
[5]marta@ufu.br

Para lidar com o desequilíbrio de classes no conjunto de dados, adotamos uma abordagem de subamostragem que selecionou aleatoriamente uma quantidade adequada de amostras da classe majoritária (0) para igualar o número de amostras entre as classes (0 indicando ausência de recidiva e 1 indicando recidiva). Além disso, empregamos validação cruzada de 5 *folds* para evitar o *overfitting*, enquanto o método de *holdout* reservou 90% dos dados para treinamento e 10% para teste. A seleção dos preditores a serem utilizados na construção dos modelos de classificação foi realizada através de uma análise de características utilizando a ferramenta *Feature Selection* no MATLAB ©, disponível no *toolbox "Classification Learner"* [4]. Esta análise possibilitou a escolha dos preditores mais relevantes, que incluíram o PSA pré-operatório, o escore de Gleason, o estadiamento e três biomarcadores: *NETO2|81831, HPN|3249 e FOLH1|2346 (PSMA).*

Com os preditores definidos, os modelos otimizados de classificação vem sendo obtidos através da busca pela árvore de decisão e SVM com base nos dados de treinamento. Posteriormente, esses modelos otimizados serão testados para determinar suas respectivas acurácias, sensibilidades e especificidades. Esse processo de escolha de dados de treino e teste foi realizado de forma aleatória e repetido 50 vezes para garantir uma análise representativa da realidade dos dados. Isso permitirá a comparação da performance dos modelos utilizando diferentes combinações de preditores, a fim de investigar se a inclusão desses biomarcadores é capaz de aprimorar a previsão da recidiva do CaP.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - código de financiamento 001.

Referências

[1] F. Alharbi e A. Vakanski. "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review". Em: **Bioengineering** 10.2 (2023), pp. 1–26. DOI: 10.3390/bioengineering10020173.

[2] Instituto Nacional do Câncer. **Câncer de Próstata: versão para profisisonais da saúde**. Online. Acessado em 11/03/2024, https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/prostata/versao-para-profissionais-de-saude. 2022.

[3] Instituto Nacional do Câncer. **Novembro Azul 2023**. Online. Acessado em 11/03/2024, https://www.gov.br/inca/pt-br/assuntos/campanhas/2023/novembro-azul. 2023.

[4] MathWorks. **Classification Learner**. Online. Acessado em 11/03/2024, https://www.mathworks.com/help/stats/classificationlearner-app.html. 2023.

[5] I. E. Naqa e M.J. Murphy. "What is machine learning?" Em: **Machine Learning in Radiation Oncology**. Ed. por I. E. Naqa, R. Li e M.J. Murphy. Springer, 2015, pp. 3–11. DOI: 10.1007/978-3-319-18305-3_1.

[6] M. Sideris e S. Papagrigoriadis. "Molecular Biomarkers and Classification Models in the Evaluation of the Prognosis of Colorectal Cancer". Em: **Anticancer Research** (2014), pp. 2061–2068. DOI: 10.1200/EDBK_351033..

[7] N. I. Simon, C. Parker, T.A. Hope e C.J. Paller. "Best Approaches and Updates for Prostate Cancer Biochemical Recurrence". Em: **American Society of Clinical Oncology Educational Book** 42 (2022), pp. 1–8. DOI: 10.1200/EDBK_351033..