1st SouthStat Meeting
Encontro Sul Brasileiro de Estatística e Ciência de Dados
14 E 15 DE DEZEMBRO DE 2023. CURITIBA-PR.

Classificação de genes associados ao câncer de mama utilizando métodos de *machine learning*

Glaucia Maria Bressan[1]
Universidade Tecnológica Federal do Paraná. Departamento Acadêmico de Matemática. Programa de Pós-Graduação em Bioinformática. Cornélio Procópio. PR
Ana Beatriz Miranda Valentin[2]
Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Bioinformática. Cornélio Procópio. PR
Leonardo Canuto Junior[3]
Universidade Tecnológica Federal do Paraná. Departamento Acadêmico de Matemática. Cornélio Procópio. PR
Elisângela Ap. da Silva Lizzi[4]
Universidade Tecnológica Federal do Paraná. Departamento Acadêmico de Matemática. Programa de Pós-Graduação em Bioinformática. Cornélio Procópio. PR

RESUMO

Introdução: Um dos problemas mais desafiadores na área de Bioinformática é o que envolve a análise de dados de expressão gênica associados a um tipo de doença, como por exemplo, o câncer de mama. Este problema envolve a interpretação de informações genéticas complexas, muitas vezes provenientes de milhares de genes. Neste contexto, o objetivo deste trabalho é desenvolver uma classificação do tipo multiclasse de genes associados ao câncer de mama, utilizando dados de expressão genética e matriz de contagem. Métodos: A base de dados é adquirida do repositório presente no chamado TCGA (*The Cancer Genome Atlas*), que consiste de um projeto colaborativo que se concentra na caracterização abrangente das alterações genômicas associadas a diversos tipos de câncer. Os dados gerados pelo TCGA são disponibilizados publicamente para a comunidade científica, permitindo que pesquisadores de todo o mundo utilizem esses dados como fonte de informações para avanços nas pesquisas de diversas áreas. Nessa base de dados foram aplicadas técnicas de pré-processamento para tratamento de dados faltantes, colunas nulas, reorganização e reestruturação dos dados, para garantir a confiabilidade das informações utilizadas nesta investigação e possibilitar o seu uso pelos algoritmos de classificação.

Em seguida, devido à dimensionalidade da base de dados (935 linhas e 14.408 colunas), foi necessário aplicar uma técnica estatística de redução da dimensionalidade, utilizando a Análise de Componentes Principais (PCA). Esta técnica é utilizada para reduzir a complexidade de conjuntos de dados multivariados, resumindo-os em um conjunto menor de componentes principais. Esses componentes capturam a maior parte da variabilidade dos dados originais, melhorarando o desempenho das tarefas de modelagem estatística e de *machine learning*, realizadas a seguir. Após a aplicação da técnica PCA, a base de dados obtida contém 935 linhas e 184 colunas. Na fase de modelagem e classificação dos genes, métodos de *machine learning* são aplicados, como Regressão Logística, *Support Vector Machine* (SVM) e *Random Forest*. Trata-se de uma classificação do tipo multiclasse, pois a base de dados apresenta 5 classes de saída: " Basal", "Her2", "LumA", "LumB", e "Normal". A comparação do desempenho dos algoritmos de classificação foi realizada por meio de indicadores estatísticos, como acurácia, matriz de confusão, F1 score e AUC-ROC. Resultados: Avaliar a eficácia dos modelos de classificação é essencial para garantir sua capacidade de generalização e aplicabilidade. Considerando a base de teste de 20% dos dados, o método de Regressão Logística obteve uma acurácia de 69%, enquanto que o SVM alcançou 75% e a *Random Forest*, 68%. Em todos os casos, a AUC-ROC e o F1-socore apresentaram valores acima de 0,85 e as matrizes de confusão contém os maiores valores na diagonal principal, indicando que os classificadores estão distinguindo bem entre as 5 classes de genes associados ao câncer de mama. Conclusão: Portanto, a abordagem proposta neste trabalho e os resultados obtidos são fundamentais para avançar em direção a tratamentos mais personalizados e eficazes para o câncer de mama, além de indicar a identificação de genes e proporcionar a correta classificação, contribuindo assim para o diagnóstico da doença.

Palavras-chave: expressão gênica; câncer de mama; Bioinformática; *machine learning*; dimensionalidade.