MODELOS DE APRENDIZADO DE MÁQUINA SUPERVISIONADO BASEADOS EM ÁRVORES APLICADOS A GENOTIPAGEM

Palavras-Chave: Machine Learning, Árvore de Decisão, Floresta Aleatória

Autores(as):

Nathan Augusto Elias, IMECC - Unicamp

Profª Drª Tatiana Benaglia, IMECC - Unicamp

INTRODUÇÃO:

O objetivo deste projeto é desenvolver conhecimento técnico-científico em Aprendizado Estatístico de Máquina aplicado à Medicina de Precisão no Brasil, além de criar ferramentas analíticas para o Projeto Jaguar, cujo foco é analisar a paisagem genômica do sistema imunológico de populações latino-americanas. Em particular, serão usados dados de scRNA-Seq que revelam a heterogeneidade celular, identificando subpopulações, estados diferenciados e padrões dinâmicos de expressão gênica, sendo crucial para entender processos como desenvolvimento embrionário, resposta imune e progressão de certas doenças.

Diante da complexidade e do volume dos dados gerados pela tecnologia de scRNA-Seq, o aprendizado de máquina torna-se uma ferramenta muito útil para classificar tipos celulares, prever estados celulares e decifrar redes de regulação gênica. Neste estudo, foram utilizados modelos baseados em árvores, como árvore de decisão e floresta aleatória. O primeiro método se baseia em criar um modelo de decisão em forma de árvore para prever uma determinada classe ou valor de saída a partir de uma entrada. O algoritmo divide iterativamente o conjunto de dados de treinamento em subconjuntos menores com base em critérios de separação para encontrar as melhores divisões que minimizam a impureza e maximizam a homogeneidade das classes em cada ramo da árvore. A combinação de muitas dessas árvores constitui o método de floresta aleatória, que busca solucionar certos problemas que uma árvore única pode possuir em cenários específicos.

METODOLOGIA:

Para realizar a pesquisa, foi usado o *software* estatístico *R Studio*, utilizando os pacotes *Bioconductor* e *Seurat* para a análise dos dados, o pacote *Tidymodels* para a aplicação dos modelos baseados em árvore e o pacote *Tidyverse* para a visualização gráfica dos dados.

O banco de dados escolhido foi um no formato *Seurat* (estrutura de dados usada na análise de scRNA-seq), que consiste em uma amostra aleatória com 20 mil regiões no transcriptoma de um banco maior, com 20 mil células e 33.234 genes de pessoas que contraíram Covid-19 e o grau de severidade: *None* (Não Detectável), *Asymptomatic* (Assintomático), *Mild* (Fraco), *Moderate* (Moderado) e *Severe* (Severo). De início, foi feita uma preparação dos dados, realizando uma limpeza a partir da filtragem de células e genes, uma normalização para remover efeitos de profundidade de sequenciamento e uma correção de *batch effects*.

Consoante ao fato de o banco de dados ser muito volumoso, foi aplicada a técnica de redução de dimensionalidade de Análise de Componentes Principais (PCA), que consiste em identificar combinações lineares a partir das variáveis do estudo e tentar reduzir a dimensionalidade escolhendo um número dessas combinações que seja menor que a dimensão original dos dados, mas que expliquem coletivamente a maioria da variabilidade no conjunto original. O primeiro eixo (ou "componente principal", PC) é escolhido de forma a maximizar essa variância, e captura os fatores dominantes de heterogeneidade no conjunto de dados. O próximo PC é escolhido de forma que seja ortogonal ao primeiro e capture a maior quantidade restante de variação, e assim por diante.

Apesar de o PCA ter sido muito útil para remover redundâncias e ruídos dos dados, ele lineariza as relações nos dados. Logo, para reverter isso, foi utilizada outra técnica de redução de dimensionalidade: a Aproximação e Projeção do Coletor Uniforme (UMAP). Ao realizar tais técnicas em sequência, foi possível obter uma visualização mais clara e informativa dos dados, capturar estruturas não lineares e aumentar a eficiência computacional, uma vez que a aplicação direta do UMAP em dados de alta dimensão pode ser computacionalmente intensa. As diferenças de visualização podem ser vistas na Figura 1.

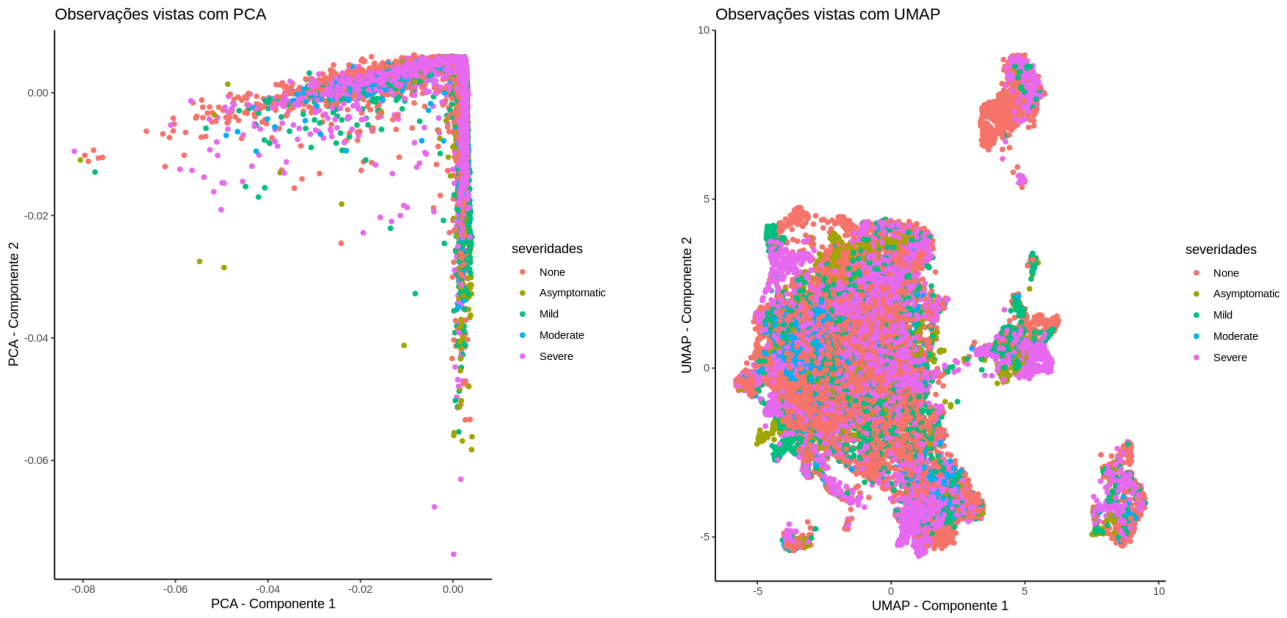

Figura 1: Distribuição das Amostras no Espaço Reduzido pelo PCA e pelo UMAP

Após a redução de dimensionalidade, foram aplicados os modelos de árvore de decisão e de floresta aleatória para analisar os dados de scRNA-Seq.

O modelo de árvore de decisão foi utilizado para criar um modelo interpretável que classifica as células com base em características específicas, permitindo uma compreensão clara das regras de decisão e das variáveis mais influentes. Isso ajuda a identificar quais características são mais importantes para a classificação das diferentes severidades da Covid-19.

Em seguida, a floresta aleatória foi empregada para melhorar a precisão e a robustez do modelo. A floresta aleatória combina múltiplas árvores de decisão, o que ajuda a reduzir o risco de overfitting e melhora a capacidade de generalização. Esse método é particularmente útil para lidar com a variabilidade e complexidade dos dados de alta dimensão, fornecendo uma análise mais estável e confiável das relações entre as características dos dados e a severidade da doença.

Para garantir a validade dos resultados e evitar viés, foi utilizada a técnica de validação cruzada (*cross-validation*). Este processo envolveu a divisão dos dados em múltiplos subconjuntos onde os modelos foram treinados em uma parte dos dados e testados em outra. Esse método permite avaliar o desempenho dos modelos de forma mais robusta, garantindo que as métricas de avaliação sejam representativas da capacidade geral dos modelos em generalizar para novos dados.

Ambos os modelos foram avaliados com base em métricas de desempenho, como acurácia, sensitividade, especificidade e a área sob a curva ROC (AUC-ROC), para determinar a eficácia na classificação das células e identificar padrões significativos associados à severidade da Covid-19.

RESULTADOS E DISCUSSÃO:

Para ambos os modelos, foram utilizados 14998 dados para o treino e 5002 para os testes. Os resultados obtidos pelo modelo de árvore de decisão podem ser vistos na Tabela 1:

Tabela 1: Matriz de confusão do Modelo de Árvore de Decisão - Preditos x Verdadeiros

	None	Asymptomatic	Mild	Moderate	Severe
None	2295	61	396	110	700
Asymptomatic	2	52	14	0	17
Mild	82	25	314	11	34
Moderate	22	0	15	31	30
Severe	245	16	99	70	361

Com a aplicação desse modelo, foram obtidos os seguintes resultados: acurácia de 0,610, sensibilidade de 0,407 e especificidade de 0,859. Esses valores indicam que, embora o modelo

apresente um desempenho razoável, com uma boa capacidade de identificar casos negativos (especificidade), há uma margem significativa para melhorias na detecção de casos positivos (sensibilidade).

O modelo de floresta aleatória, gerou os resultados presentes na Tabela 2:

Tabela 2: Matriz de confusão do Modelo de Floresta Aleatória - Preditos x Verdadeiros

	None	Asymptomatic	Mild	Moderate	Severe
None	2254	50	319	95	611
Asymptomatic	4	56	22	0	8
Mild	92	16	389	10	62
Moderate	30	0	17	53	28
Severe	264	34	99	65	422

Da mesma forma que ocorreu com o modelo de árvore de decisão, o modelo de floresta aleatória apresentou resultados que ainda não são totalmente satisfatórios, com uma acurácia de 0,635, uma sensibilidade de 0,456 e uma especificidade de 0,872. Esses resultados indicam que, embora o modelo tenha melhorado em relação ao anterior, ainda há necessidade de aplicar métodos adicionais para aprimorar a precisão das predições.

CONCLUSÕES:

Consoante a tudo que foi apresentado, nota-se que a aplicação dos modelos baseados em árvores foi adequada devido à sua capacidade de lidar com dados não lineares, mas ainda há espaço para melhorias na predição dos dados. Abordagens complementares, como superamostragem ou subamostragem, podem ser necessárias para aprimorar a eficácia dos modelos. Portanto, enquanto esses modelos forneceram uma base sólida para a análise, eles evidenciam a necessidade de ajustes e técnicas adicionais para melhorar a precisão e a eficácia na classificação das severidades da Covid-19.

BIBLIOGRAFIA

AMEZQUITA, R. A.; LUN, A. T. L.; BECHT, E.; CAREY, V. J.; CARPP, L. N.; GEISTLINGER, L.; MARINI, F.; RUE-ALBRECHT, K.; RISSO, D.; SONESON, C.; WALDRON, L.; PAGÈS, H.; SMITH, M. L.; HUBER, W.; MORGAN, M.; GOTTARDO, R.; HICKS, S. C. **Orchestrating single-cell analysis with Bioconductor**. *Nature Methods*, v. 17, p. 137-145, 2020.

CARVALHO, B. S.; LOUIS, T. A.; IRIZARRY, R. A. **Quantifying uncertainty in genotype calls**. *Bioinformatics*, v. 26, p. 242-249, 2010.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. ***An Introduction to Statistical Learning: with Applications in R***. New York: Springer, 2013.

GENTLEMAN, R.; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ... & GATTO, L. **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology*, v. 5, R80, 2004.

SATIJA, R.; FARRELL, J. A.; GENNERT, D.; SCHIER, A. F.; REGEV, A. **Spatial reconstruction of single-cell gene expression data**. *Nature Biotechnology*, v. 33, p. 495-502, 2015.

SCHARPF, R. B.; IRIZARRY, R. A.; RITCHIE, M. E.; CARVALHO, B.; RUCZINSKI, I. **Using the R Package crlmm for Genotyping and Copy Number Estimation**. *Journal of Statistical Software*, v. 40, p. 1-32, 2011.

WICKHAM, H.; GROLEMUND, G. ***R for Data Science: Import, Tidy, Transform, Visualize, and Model Data***. O'Reilly Media, Inc., 2017.

ZHANG, Z. H.; JHAVERI, D. J.; MARSHALL, V. M.; BAUER, D. C.; EDSON, J.; NARAYANAN, R. K.; ROBINSON, G. J.; LUNDBERG, A. E.; BARTLETT, P. F.; WRAY, N. R.; ZHAO, Q. Y. **A comparative study of techniques for differential expression analysis on RNA-Seq data**. *PLoS One*, v. 9, e103207, 2014.

KUHN, M.; WICKHAM, H. *Tidymodels:* ***A Collection of Packages for Modeling and Machine Learning using Tidyverse Principles***. 2020. Disponível em: https://www.tidymodels.org.