Classificação de Grupos Celulares via Aprendizado de Máquina

Palavras-Chave: Aprendizado Estatístico de Máquina, scRNA-Seq, K-vizinhos mais próximos

Autores(as):

Ana Julia Cunha e Silva, IMECC – UNICAMP

Prof. Dr. Benilton de Sá Carvalho, IMECC - UNICAMP

INTRODUÇÃO:

O objetivo deste projeto é estudar a técnica supervisionada KNN (K-vizinhos mais próximos) dentro de um contexto de aprendizado estatístico de máquina. Existem diversas aplicações do método, mas nesse projeto foi utilizado dados de severidade de Covid-19. Os dados usam a tecnologia de scRNA-Seq (*single-cell RNA-sequencing*).

Os dados de scRNA-Seq são volumosos e complexos. Técnicas de aprendizado estatístico de máquina podem nos ajudar a compreender com maiores detalhes as interações que acontecem em sistemas biológicos para a expressão gênica em um organismo complexo, como o humano. Em scRNA-Seq, os métodos de aprendizado de máquina supervisionado são utilizados para a identificação de padrões de expressão e classificação de células em grupos pré-definidos. Para a análise de dados utilizamos o ambiente R/Bioconductor.

METODOLOGIA:

Dados gerados por scRNA-seq são dados de alta-dimensão, a técnica sequência o RNA presente em cada uma das células amostradas de doadores. A análise de dados, em termos de Ciência de Dados, começa com a disponibilização da matriz de contagem, em que as linhas representam genes e as colunas representam células (que podem pertencer a um mesmo indivíduo). Cada casela desta matriz de contagem representa quantas vezes um dado gene foi observado naquela célula. Por se tratar de observações em uma unidade experimental tão "pequena", esta é uma matriz enriquecida por zeros. Por este motivo e para uso eficiente de recursos computacionais disponíveis, a matriz de contagem é frequentemente representada de forma esparsa. Sendo assim, para podermos analisar de fato o conjunto de dados, é necessária uma fase de preparação dos dados.

Esta preparação é crucial para conseguirmos responder apropriadamente perguntas de cunho biológico. A primeira etapa é a retirada dos ruídos de origem técnica para podermos explicar a variabilidade como função de fatores biológicos. A segunda etapa consiste na detecção e correção dos efeitos de lote (batch effects), assim garantimos que as análises vão refletir as diferenças biológicas e

não as distorções de origem técnica. As células são sequenciadas em diferentes lotes, as diferenças sistemáticas observadas entre esses lotes são os efeitos de lote.

Sabendo que um ser humano possui cerca de 20.000 genes por célula, a matriz correspondente aos dados possui uma dimensão proporcional à quantidade de genes analisados e células doadas. Para um estudo com N voluntários onde cada um contribui com exatamente K células, a matriz vai ter dimensão aproximada de 20.000 x N x K. Além da dimensão ficar muito grande, muitos genes têm pouca variabilidade entre si, fazendo com que a informação não seja relevante.

A matriz de contagem dos dados utilizados contém 20 mil células e pouco mais de 33 mil genes, entretanto para nos adequarmos às condições técnicas do equipamento disponível e tentar reduzir a quantidade de células originadas de um mesmo doador, seguimos a estratégia de reduzir o número de genes estudados. Foi realizada a seleção dos 2 mil genes com maior variabilidade e então seguimos para a modelagem com a técnica supervisionada KNN para classificar as severidades. O KNN é uma técnica supervisionada baseada em distância entre pontos, que pode ser usada tanto para problemas de classificação quanto de regressão.

No aprendizado de máquina dividimos os dados em dois conjuntos, o treino e o teste. O treino é importante para que o modelo aprenda os padrões dos dados, enquanto o teste serve para conseguirmos avaliar o desempenho para dados novos que não foram aprendidos ainda. O tamanho dos conjuntos podem variar entre 70% a 80% para o treino e 20% a 30% para teste, aqui a amostra de treino consiste em 14.999 observações e as outras 5.001 ficam para teste.

Dado o desbalanceamento das frequências de alguns níveis de severidade, foi necessário aplicar uma re-amostragem para que todos os níveis tivessem a mesma frequência. Definiu-se o dobro da frequência do nível minoritário, usando um misto de super-amostragem com sub-amostragem.

Para a aplicação do método é necessário definirmos alguns hiperparâmetros, sendo eles a quantidade de vizinhos, a fórmula para calcular a distância e a função kernel para a distância. A definição desses hiperparâmetros ocorreu por meio de uma validação cruzada, usando os dados da amostra de treino, com 5 folds e a escolha do melhor modelo foi para aquele com maior acurácia. Foram comparados modelos com 3 valores diferentes para a quantidade de vizinhos e 9 tipos de função de kernel.

Após encontrarmos o melhor modelo, aplicamos ele na amostra de teste e fizemos a predição. Por fim comparamos os resultados preditos com os verdadeiros na matriz de confusão, assim conseguimos ver quantos acertos e erros o modelo teve para cada severidade.

RESULTADOS E DISCUSSÃO:

O melhor modelo, dentre os 18 testados na validação cruzada, foi aquele com 10 vizinhos, distância euclidiana e função de kernel triangular. O modelo teve uma acurácia de 0,38 e a estatística Kappa foi de 0,189 quando aplicado na amostra de teste. Quanto mais próxima de 1,0 a acurácia for, menor o erro cometido na estimativa ou medição, enquanto a Kappa mede a avaliação das classificações entre categorias (quanto mais próximo de 1,0 melhor). Temos pelas estatísticas que o

modelo não parece ter bom desempenho nas predições, entretanto dada a complexidade do conjunto dos dados, o resultado é muito positivo.

Na Tabela 1 é possível ver as frequências em cada conjunto de dados. A escolha de reamostragem teve impacto direto na modelagem, nota-se que os níveis não seguem mais as proporções da base total e na amostra de treino o nível “Nenhum” tem mais observações do que foi utilizado para modelar. Entretanto como ponto positivo, diminuímos a quantidade de erros de alocação para o nível maioritário é menor.

Frequências das Severidades por Conjunto de Dados					
	Nenhuma	**Assintomático**	**Leve**	**Moderada**	**Severa**
Base Total	10591	680	3347	834	4548
Amostra de Treino	7956	504	2511	619	3409
Treino Reamostrado	1008	1008	1008	1008	1008
Amostra de Teste	2635	176	836	215	1139

Tabela 1 – Tabela de Frequências das classes das Severidades em cada base de dados trabalhada.

Na Tabela 2 vemos a matriz de confusão, nela podemos ver que os níveis que mais tiveram atribuições falsas foram “Nenhuma” e “Severo” que são os 2 mais representados nos dados. Num geral vemos pela diagonal que os níveis foram mais acertados do que errados, principalmente nos níveis menos representados (“Assintomático” e “Moderado”) na base original.

Nível Predito	**Nível Verdadeiro**				
	Nenhuma	**Assintomático**	**Leve**	**Moderado**	**Severo**
Nenhuma	801	22	84	6	146
Assintomático	162	86	33	6	54
Leve	578	28	346	30	231
Moderado	245	7	97	115	134
Severo	849	33	276	58	574

Tabela 2 – Matriz de confusão.

CONCLUSÕES:

Conforme visto nos resultados estatísticos, o modelo não conseguiu discriminar completamente os grupos de severidade, entretanto o resultado geral não é negativo. Os dados são muito complexos, possuímos apenas as expressões gênicas de várias células, onde muitas são dos mesmos doadores. Quando temos mais de uma célula proveniente de um doador, essas células serão correlacionadas entre si, possuindo a mesma resposta.

Pelo KNN ser um método baseado em distância entre pontos, é esperado um desempenho com estatísticas baixas conforme o tamanho da dimensão dos dados. Num geral o resultado foi bem

positivo, dada a complexidade dos dados e a característica do KNN, foi possível aplicar um bom modelo.

BIBLIOGRAFIA

LIAO M, LIU Y, YUAN J, WEN Y et al. **Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19.** Nat Med 2020 PMID: 32398875
medRxiv: https://doi.org/10.1101/2020.02.23.20026690 Acesso em: 08/03/2024

DERYCKEL, F. **Machine Learning with R.** 2019. Disponível em: https://fderyckel.github.io/machinelearningwithr/ Acesso em: 08/03/2024

LIAO M, LIU Y, YUAN J, WEN Y, XU G, ZHAO J, CHENG L, LI J, WANG X, WANG F, LIU L, AMIT I, ZHANG S, ZHANG Z. **Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19.** Nat Med. 2020 doi: 10.1038/s41591-020-0901-9. Epub 2020 May 12. PMID: 32398875. Acesso em 08/03/2024

KUHN M, SILGE J. **Tidy Modeling with R.** 2023. Disponível em: https://www.tmwr.org Acesso em 02/08/2024