XXXI Congresso de Iniciação Científica Unicamp 2023

Aprendizado (Supervisionado) Estatístico de Máquina Aplicado a Dados de scRNA-Seq

Palavras-Chave: Machine learning, genética, binomial negativa

Autores(as):

Mateus Costa Trentini, IMECC – Unicamp

Prof(ª). Dr(ª). Benilton de Sá Carvalho, IMECC - Unicamp

INTRODUÇÃO:

Este projeto tem como objetivo proposto o desenvolvimento na modelagem de dados de alta-dimensão de origem biológica, utilizando técnicas de análise estatística e aprendizado estatístico de máquina. Em particular, foca-se na análise de dados de single-cell RNA-seq, com um dos intuítos alimentar análises complementares conduzidas por equipes biomédicas na caracterização do sistema imunológico de indivíduos latino-americanos.

Um dos principais focos foram os ensaios de sequenciamento genético de alta capacidade comparativos entre diferentes grupos. Para isso uma tarefa fundamental foi a análise de dados de contagem, no caso as contagens de leitura por gene em RNA-seq, com objetivo de obter evidências de mudanças sistemáticas em condições experimentais. Poucas repetições, o fato das contagens serem discretas, com amplo intervalo, alta variância e outliers requerem uma abordagem estatística adequada. Assim, usamos o DESeq2, um método para análise diferencial de dados de contagem, usando shrinkage e mudanças de proporção nas variáveis para melhorar a estabilidade e interpretabilidade das estimativas.

Para estimar as expressões diferenciais entre grupos precisamos fazer comparações entre as diferentes contagens de genes entre os tratamentos e, para tal, se faz necessário usar um modelo probabilístico apropriado para os dados, dessa forma apresentamos a distribuição Binomial Negativa e as vantagens de a usar nesse contexto. Além disso fizemos comparações entre grupos para um conjunto de dados experimentais usando o DESeq2 e em seguida usaremos técnicas de aprendizado de máquina para construir modelos com estes dados fazendo uso do tidymodels e outros diversos pacotes.

METODOLOGIA:

Como citado, o principal objetivo da pesquisa é a modelagem de dados RNA-seq e scRNA-seq, e, para isso, é necessário o uso de um modelo apropriado para as contagens dos genes, estamos lidando com dados discretos e, ao considerar a leitura de um determinado gene entre todos os outros, com probabilidade baixa de acontecer, assim durante a pesquisa foi demonstrada a razão e efetividade do uso da binomial negativa como distribuição para a modelagem destas contagens.

Com a distribuição para os dados definida passamos a exemplos e atividades práticas, primeiro conjunto de dados estudado foi o do pacote `pasilla` no R, no entanto o maior foco foi no segundo conjunto sobre o qual discutiremos a respeito: Os dados experimentais do pacote `FieldEffectCrC`, disponível entre os conjuntos de dados experimentais no Bioconductor. Estes são dados processados RNA-seq de 1139 amostras de tecido colorretal primário humano em três fenótipos, incluindo tumores,

normais adjacentes aos tumores e saudáveis. Estes dados foram adquiridos ao harmonizar e sumarizar todas as amostras públicas deste tipo de dado ao longo dos anos.

Carregamos e processamos os dados obtendo um objeto da classe DESeqDataSet, com 37361 linhas, os genes, e 834 colunas, as amostras. Alguns genes tem contagens baixas demais, sugerindo que, na realidade, estes genes não estavam presentes na amostra biológica coletada, de forma que observamos apenas ruído. Sendo assim, optamos por remover genes que foram observados menos que dez vezes entre todas as amostras.

Fizemos também o shrinkage dos dados para lidar com outros problemas das diferenças de contagem, incluindo efeitos sistemáticos observados na ocorrência de variâncias muito baixas. Depois dos tratamentos mencionados, podemos visualizar as expressões diferenciais entre tecidos com câncer e saudáveis:

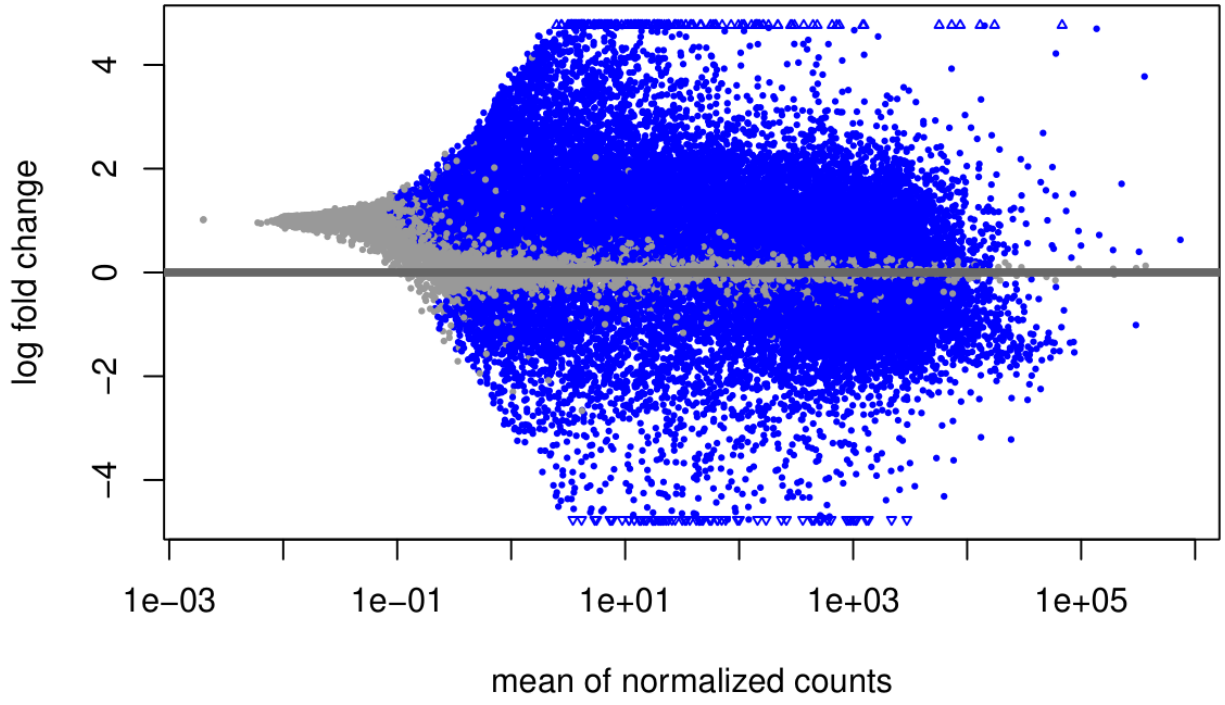

Figura 1Visualização das expressões diferenciais entre tecido com câncer e saudável.

Na outra parte do projeto, em uma outra visão, podemos estar interessados em identificar a qual grupo a pessoa pertence dadas as contagens dos seus genes, contrário a apenas verificar as diferenças das contagens dados os grupos, e esse estudo é precisamente uma aplicação de Aprendizado Supervisionado de Máquina, onde nossas variáveis preditoras são as contagens de cada gene e a variável resposta é classificação do grupo de determinado tecido.

Para esse estudo lidamos apenas com os grupos HLT (saudável) e CRC (câncer), assim, obtemos os dados utilizando alguns dos métodos disponíveis no DESeq2 como fizemos nas análises anteriores e lidamos com os dados pre-processados, já que o pacote tem normalizações e shrinkage específicos para este tipo de dado. Em seguida, processamos os dados usando métodos contidos no `tidymodels`, pacote que engloba diversos outros pacotes de processamento e modelagem de dados.

Para obtenção dos dados, é útil usar alguns dos métodos disponíveis no DESeq2 como fizemos anteriormente e lidar com os dados pre-processados, já que o pacote tem normalizações e shrinkage específicos para este tipo de dado. Assim transformamos os dados em um dataframe e fazemos os primeiros processos para a modelagem.

Em seguida, construimos um modelo de `random forest`, com argumentos como o mínimo de observações por nó final e número de variáveis selecionadas em cada nó como hiper-parâmetros a serem otimizados. Fazemos a otimização desses hiper-parâmetros usando validação cruzada, então, com o modelo ajustado e otimizado, podemos tirar as métricas como acurácia do modelo nos dados de teste, que não usou para treino. Resultando em acurácia e área sob a curva (AUC) de aproximadamente 99%, consideradas sem dúvidas muito altas. Assim, vemos que o modelo funcionou bem tanto para os dados de treino quanto de teste.

RESULTADOS E DISCUSSÃO:

Como vimos, a modelagem resultou em predições com uma performance muito alta, indicando que o modelo foi apropriado para o problema e que, ao menos para estes dados, Aprendizado de Máquina foi uma ótima abordagem para as questões que levantamos, além disso, do modelo foi possível extrair as importâncias das variáveis, os genes, encontrando de uma forma alternativa ao DESeq2 genes de maior importância na separabilidade dos grupos.

Nas análises vimos também que há muitos genes com expressões diferenciais significativamente diferentes, o que também indicou a separabilidade dos grupos e, assim, uma possível boa performance de modelos de Aprendizado de Máquina, resultando na corroboração das análises.

CONCLUSÕES:

Observamos os desafios da modelagem de dados genéticos e como a Distribuição Binomial Negativa trata corretamente particularidades observadas em sistemas biológicos. Os resultados obtidos para os dados de tecidos com câncer e saudáveis mostram que há diferenças nas contagens entre os grupos analisados e que a modelagem preditiva é possível e útil.

Os resultados do estudo trazem otimismo para o avanço das técnicas de Aprendizado de Máquina no contexto genético e biológico. Há ainda muitas aplicações para os modelos já explorados e muitos possíveis modelos para os dados que foram usados, havendo assim muito espaço para aprendizado e pesquisa.

Os aprendizados e técnicas desenvolvidas durante a pesquisa servirão também de auxílio para novos pesquisadores aprendendo tanto sobre os dados específicos do domínio quanto para interessados por Aprendizado de Máquina.

BIBLIOGRAFIA

Love, M.I., Huber, W., e Anders, S. (2020). **FieldEffectCrc: RNAseq dataset for colon carcinoma cell lines with 5-azacytidine treatment.**

https://bioconductor.org/packages/release/data/experiment/html/FieldEffectCrc.html. Acesso em: 01 de março de 2023.

Love, M.I., Huber W, Anders S. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** Genome Biology. 2014;15(12):550. Disponível em:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

Love, M.I., Anders, S., e Huber, W. (2023). **DESeq2: Differential gene expression analysis based on the negative binomial distribution.**

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/ inst/doc/DESeq2.html. Acesso em: 01 de março de 2023.
TidyModels. (2023). Recuperado em 01 de março de 2023, de https://www.tidymodels.org
Kuhn, Max, and Silge, Julia. Tidy Modeling with R. N.p., O'Reilly Media, 2022.
Anders, Simon, and Wolfgang Huber. 2010. **"Differential Expression Analysis for Sequence Count Data."** Genome Biology 11: R106. http://genomebiology.com/2010/11/10/R106. Recuperado em 01 de março de 2023.