Abordagens de aprendizado de máquina não supervisionado aplicadas a dados de single-cell RNA-Seq (scRNA-seq)

Palavras-Chave: Aprendizado Estatístico de Máquina; scRNA-Seq e Modelos de Classificação.

Autores(as):

HELOISA DE OLIVEIRA GARCIA – UNICAMP

Prof(ª). Dr(ª). BENILTON DE SÁ CARVALHO (orientador) – UNICAMP

INTRODUÇÃO:

Este projeto integra uma iniciativa de coleta de amostras de sangue de indivíduos latino-americanos participantes da rede Human Cell Atlas (HCA). Utilizando a técnica de single-cell RNA-Seq (scRNA-seq), o projeto visa analisar a expressão gênica a nível celular e identificar variantes genéticas. Essa abordagem permite a detecção de subgrupos de indivíduos com características genéticas compartilhadas e respostas específicas a fenótipos como doenças.

Este estudo, que será apresentado no XXXII Congresso de Iniciação Científica da Unicamp, investiga por que a COVID-19 é geralmente mais branda em crianças comparado aos adultos, analisando pacientes pediátricos e adultos com COVID-19 e indivíduos saudáveis através de modelos de classificação.

METODOLOGIA:

Com o objetivo de caracterizar a paisagem genômica do sistema imunológico em populações latino-americanas utilizando técnicas de sequenciamento de RNA, especialmente o sequenciamento de célula única, o projeto lida com dados volumosos e de alta dimensionalidade. O foco principal é identificar determinantes genéticos da variabilidade do sistema imunológico, compreendendo seus efeitos em diferentes contextos celulares.

Os dados estão publicamente disponíveis por meio da iniciativa Human Cell Atlas (HCA) e representam a contagem de cada gene observado em uma dada célula e existe uma hierarquia entre células e seus doadores. Desta maneira, a representação genérica destes dados se dá por meio de uma matriz com R linhas (20.000 genes identificados no genoma humano) e C colunas (onde cada coluna representa uma célula que podem ser agrupadas de acordo com seus doadores).

Em dados de altas dimensões, a tarefa de representar os dados graficamente torna-se cada vez mais complexa e menos intuitiva. A fim de contornar este problema, utiliza-se a PCA, que busca identificar combinações lineares escolhendo um número destas combinações que seja menor que a dimensão original dos dados, mas que expliquem coletivamente a maioria da variabilidade no conjunto original (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2013). A componente principal é (PC) é escolhida de forma a maximizar essa variância, e capturar os fatores dominantes de heterogeneidade no conjunto de dados. A próxima PC é escolhida de forma que seja ortogonal a primeira e

capture a maior quantidade restante de variação, e assim por diante.

Essa estratégia da utilização da PCA vem sendo combinada com a UMAP, proposta por McInnes et al. (2018), como um algoritmo de redução de dimensionalidade não linear que, de forma sintetizada, constrói uma representação gráfica de alta dimensão dos dados e, em seguida, otimiza um gráfico de baixa dimensão para ser o mais estruturalmente semelhante possível. Para o primeiro, ele constrói uma representação de um gráfico ponderado, com pesos de aresta representando a probabilidade de dois pontos estarem conectados. Para determinar a conectividade, ele estende um raio para fora de cada ponto, conectando pontos quando esses raios se sobrepõem. Ele escolhe o raio localmente, com base na distância até o n-ésimo vizinho mais próximo de cada ponto e então torna o gráfico difuso, ou seja, que se espalha largamente por todas as direções diminuindo a probabilidade de conexão à medida que o raio aumenta. Finalmente, ao estipular que cada ponto deve estar conectado pelo menos ao seu vizinho mais próximo, a UMAP garante que a estrutura local seja preservada em equilíbrio com a estrutura global.

No aprendizado não supervisionado, apenas o conjunto de dados é fornecido e a principal tarefa é formar uma partição que divide os dados em grupos de objetos similares, maximizando a homogeneidade entre os objetos de um mesmo grupo, da mesma forma que maximize a heterogeneidade entre os objetos de grupos distintos. Pode-se classificar uma partição como sendo rígida ou soft (Anderson et. al., 2010). Na primeira, cada objeto pertence a um único grupo. Já na segunda, a binarização em relação a um objeto pertencer a um grupo é flexível, de forma que seja útil em problemas nos quais existe sobreposição de grupos. É importante ressaltar que a maioria dos algoritmos descritos na literatura assume que o número de grupos é fornecido pelo usuário. Dessa forma, se concentram em obter K grupos de objetos semelhantes de acordo com algum critério pré estabelecido.

Um Gaussian Mixture Model (GMM), ou modelo de misturas gaussianas é um modelo estocástico, que utiliza distribuições probabilísticas para modelar os dados, estima os parâmetros dos modelos de forma iterativa usando o algoritmo de Expectation Maximization (EM), e considera a incerteza na associação dos dados aos clusters. A abordagem assume que os dados consistem em uma mistura de distribuições, cada uma representando um cluster distinto. Ao estimar os parâmetros desses componentes, o clustering GMM identifica e separa pontos de dados pertencentes a diferentes clusters. Quando usamos modelos gaussianos, cada componente assume uma distribuição normal multivariada.

O uso de EM pode ser visto, do ponto de vista de agrupamentos de dados, como o particionamento dos dados em K grupos de forma probabilística. Portanto, cada componente representa um grupo. Diferentemente do particionamento rígido dos dados, neste modelo cada objeto pertence a um grupo com certa probabilidade. Caso se considere cada objeto pertencendo ao grupo de maior probabilidade e matrizes de covariância diagonais e proporcionais a matriz de identidade, o EM para GMM se reduz ao algoritmo K-Means.

Nas últimas décadas tem sido estudado o uso de Algoritmos Evolutivos (AEs) (Eiben e Smith, 2003) em problemas de agrupamentos de dados. Eles consistem em algoritmos de busca estocásticos, e baseiam-se no processamento de um conjunto de soluções de tal forma a resolver um determinado problema através da otimização matemática. Em suma, são capazes de obter boas soluções para problemas difíceis em tempo razoável. Uma das vantagens é sua robustez a ótimos locais em relação aos métodos de busca tradicionais.

Naldi et al (2011) realizaram diversos experimentos empíricos que indicam que AEs bem projetados podem ser eficientes computacionalmente para otimizar partições obtidas

do algoritmo K-means. Apesar do EM para GMM ser amplamente utilizado, e diversas variantes terem sido desenvolvidas, poucas AEs foram propostas para este problema.

RESULTADOS:

Os dados tem formato de uma matriz esparsa, ou seja, com a maioria dos elementos iguais a zero e uma amostra disso pode ser visualizada na Figura 1.

	AAACCTGTCGTAGGAG-CV001_KM8853544-ucl	AAAGCAATCATCGCTC-CV001_KM8853544-ucl	AACCGCGAGCACGCCT-CV001_KM8853544-ucl	AACTGGTGTAATAGCA-CV001_KM8853544-ucl	AAGACCTCAATGGTCT-CV001_KM8853544-ucl
ENSG00000121410	0	0	0	0	0
ENSG00000268895	0	0	0	0	0
ENSG00000148584	0	0	0	0	0
ENSG00000175899	0	0	0	0	0
ENSG00000245105	0	0	0	0	0

Figura 1: Amostra dos dados com elementos iguais a 0.

Estamos interessados nos genes com as maiores pontuações de variabilidade, portanto ao atualizar os dados seguindo esse critério, removemos genes com baixa variação de expressão. Como o conjunto de dados é muito grande, aplicamos uma normalização e em seguida a PCA para reduzir a dimensionalidade. Para cada componente principal, os genes que contribuem positivamente e negativamente para a variabilidade desta componente são listados. Esses genes podem ser interpretados como os principais drivers da variabilidade observada nos dados. Em sequência, capturamos a estrutura subjacente dos dados de expressão gênica em um espaço de menor dimensão, facilitando a visualização e a análise dos padrões de expressão gênica através do UMAP.

A Figura 1 representa as coordenadas UMAP das células nas cinco primeiras componentes, mas ao invés de analisar todas elas, escolhemos somente as duas dimensões principais (Figura 1 - Gráfico 1). Para enriquecer a análise, adicionamos às coordenadas de UMAP, uma coluna chamada ‘sev’, que representa níveis de gravidade da COVID-19. Podemos avaliar como os dados se comportam através de um gráfico de dispersão das células, colorindo-o de acordo com a gravidade da COVID-19 (Figura 1 - Gráfico 2).

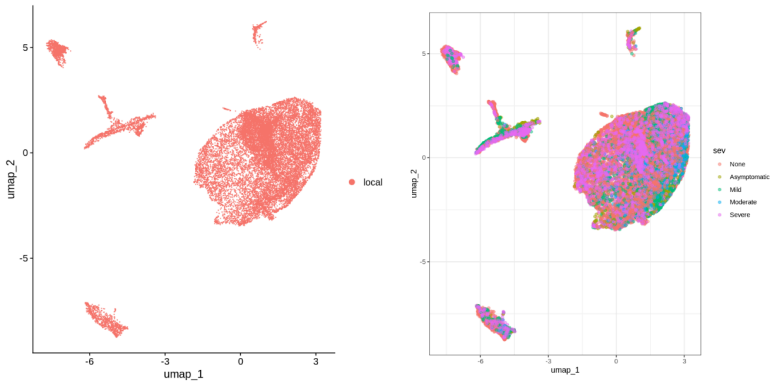

Figura 1: Gráfico 1: UMAP aplicado a duas dimensões. Gráfico 2: UMAP com adicional dos níveis de gravidade da COVID-19.

Não conseguimos distinguir como os graus de severidade da doença estão se distribuindo pelas células, e como podemos agrupá-los seguindo um padrão. A partir disso, aplicamos o Agrupamento K-Means. A fim de ser mais assertivos e definir a quantidade ideal de clusters (k) para iniciar a iteração, construímos uma função que calcula a soma dos quadrados intra-cluster para diferentes valores de k usando o algoritmo k-means, e podemos visualizá-la através de um Elbow Plot (Figura 2 - Gráfico 1).

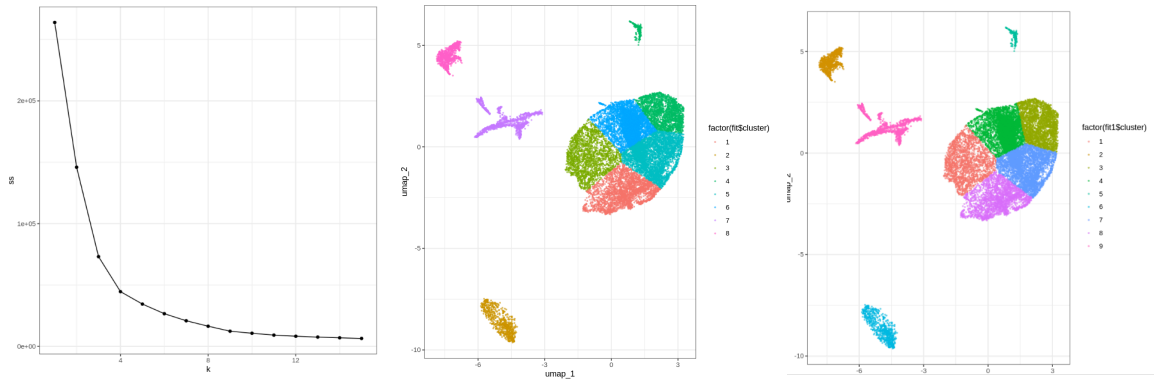

Figura 2: A esquerda, o Elbow Plot, representando a quantidade razoável de clusters. Ao centro e à direita, a aplicação do algoritmo k-means para k=8 e k=9, respectivamente.

Ao substituir a análise para o GMM criamos um algoritmo que executa uma repetição, para cada valor de G (número de clusters), ajustando o modelo aos dados. Em seguida, as atribuições de cluster são obtidas, e os pontos são plotados em um gráfico de dispersão bidimensional, onde cada ponto é colorido de acordo com o cluster ao qual foi atribuído. Isso permite comparar visualmente os resultados do agrupamento para diferentes números de clusters e avaliar qual número de clusters parece mais apropriado para os dados em questão, com base na estrutura identificada nos gráficos.

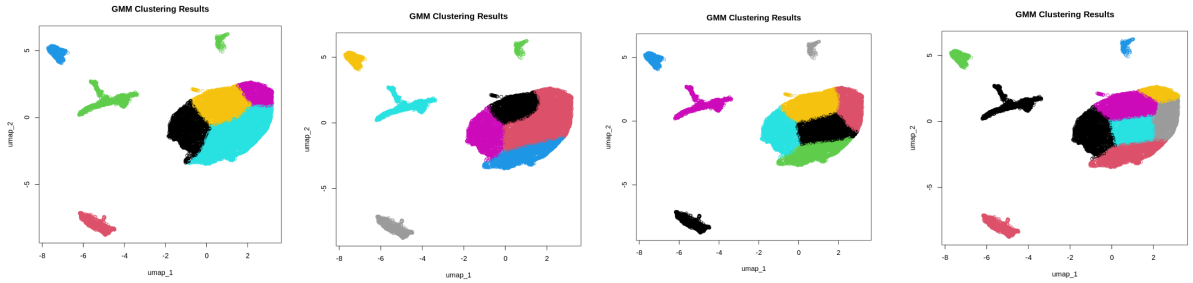

Figura 3: Clustering GMM aplicado aos dados com o número de grupos variando de 7 a 10.

O GMM fornece uma probabilidade de pertencimento de cada ponto de dados a cada cluster, o que pode ser usado para estimar a incerteza nas atribuições de cluster. As células têm diferentes probabilidades de pertencer a diferentes clusters, indicando que o modelo GMM consegue diferenciar bem entre os diferentes grupos de células.

A célula pode ser atribuída ao cluster com a maior probabilidade. Isso é o que geralmente se faz para fins de classificação. Células com probabilidades semelhantes para dois ou mais clusters podem ser analisadas em termos de suas características biológicas, pois podem representar estados transicionais ou populações mistas.

Cada cluster em um GMM é caracterizado por uma média (centro) e uma variância (dispersão) das suas distribuições. Alta variância indica que os pontos estão espalhados amplamente em torno da média. Baixa variância indica que os pontos estão mais próximos do centro. As elipses de confiança, como mostrado na Figura 4 (Gráfico 2 e 4), ajudam a visualizar a distribuição e a dispersão dos pontos dentro de cada cluster. Este gráfico permite visualizar como os pontos foram agrupados pelo modelo GMM, com a cor indicando o cluster e a opacidade indicando a confiança na atribuição de cluster.

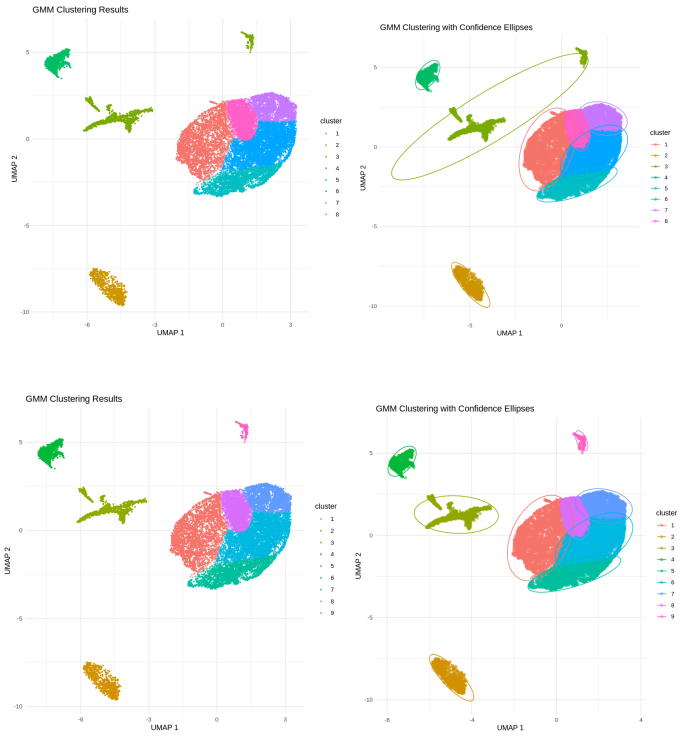

Figura 4: Agrupamento por Modelo de Misturas Gaussianas (GMM), com elipses de confiança para 8 e 9 grupos,.

CONCLUSÃO:

Podemos concluir que tanto PCA quanto UMAP facilitam a forma de trabalhar com os dados. Em relação ao Clustering, o algoritmo GMM teve uma performance satisfatória, assim como será ainda mais rico continuar investido em variantes dos Algoritmos Evolutivos. Será necessário um profissional da área RNA-seq para que estudem como as amostras combinadas permitem estudar a dinâmica espacial da infecção, fornecendo uma interface para esses dados como uma referência detalhada para o estudo de respostas imunes a crianças.

BIBLIOGRAFIA:

Anderson, D.T., Bezdek, J.C., Popescu, M., e Keller, J.M. (2010) Comparing fuzzy, probabilistic and possibilistic partitions. IEEE Transactions on Fuzzy Systems.
Eiben. A. E. Smith, J. E. (2003). Introduction to Evolutionary Computing. Springer.
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer (2013).
Hongyu, et al, E&S - Engineering and Science. 5:1; (2015).
JOHNSON, R.A.; WICHERN, D.W. Applied multivariate statistical analysis.Madison: Prentice Hall International. 816p; (1998).
Lloyd, S., Least squares quantization in PCM, IEEE transactions on information theory 28.2: 129-137, 1982
MCINNES, L. et al. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, v. 3, n. 29, p. 861, set. 2018. ISSN 2475-9066. Disponível em: http://joss.theoj.org/papers/10.21105/joss.00861.
Naldi, M. C., Campello, R. J.Hruschka, E. R., e Carvalho. A. C (2011). Efficiency issues evolutionary k-means. Applied Soft Computing 11 (2): 1938-1952.