SOM+K-means *vs.* K-means para Agrupamento de Imagens de Células de Leucemia Linfoblástica

Vinicius Rezende Bardelin[1]

[1]Instituto de Matemática, Estatística e Computação Científica – UNICAMP

1. Introdução

A detecção precisa de células cancerígenas de leucemia desempenha um papel crucial na saúde e na qualidade de vida, fornecendo informações para diagnósticos precoces e estratégias de tratamento. Para alcançar esse objetivo, técnicas avançadas de análise de dados podem ser utilizadas, permitindo a identificação de subtipos celulares com características distintas e o entendimento da progressão da doença.

A leucemia linfoblástica aguda (LLA) é um tipo de câncer sanguíneo que se desenvolve rapidamente e é caracterizada pela presença excessiva de linfoblastos, células imaturas que podem se transformar em linfócitos maduros, no sangue e na medula óssea. Esses linfoblastos se proliferam de forma descontrolada, comprometendo a produção normal de células sanguíneas e afetando o funcionamento do sistema imunológico.

Neste artigo, comparamos dois métodos de agrupamento para segmentar imagens de células saudáveis e não saudáveis do sangue: Self-Organizing Maps (SOM) combinado com K-means (denotado por S+K) e K-means (isolado). Utilizamos o dataset ALL-IDB (Acute Lymphoblastic Leukemia Image Database), que consiste em 260 imagens divididas igualmente entre células saudáveis e não saudáveis. A Figura 1 exibe algumas amostras do conjunto de dados.

2. Métodos

2.1. Self Organizing Maps (SOM)

Introduzido por Kohonen, o SOM é uma técnica de aprendizado não supervisionado pertencente a classe de redes neurais que mapeia dados multidimensionais em uma grade tipicamente bidimensional. O SOM busca organizar os dados de forma que regiões similares no espaço de entrada $\mathbf{X}$ sejam representadas por neurônios que são vizinhos na grade SOM, preservando as relações topológicas originais dos dados.

Matematicamente, o SOM é treinado para minimizar a distância entre cada amostra de entrada x e seu neurônio mais próximo μ_i, onde μ_i representa o peso do i-ésimo neurônio na grade SOM. Isso é formalizado pela seguinte função de custo:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{M} d(i,j) \|x - \mu_{ij}\|^2$$

Onde N e M são as dimensões da grade, $d(i,j)$ é a medida de proximidade entre os neurônios i e j, e μ_{ij} representa o peso do neurônio na posição (i,j) na grade SOM.

O treinamento do SOM envolve a atualização dos pesos dos neurônios com base na distância entre as amostras $\mathbf{X}$ e os neurônios na grade SOM. Os pesos são ajustados de forma iterativa para refletir as características dos dados de entrada, resultando em uma organização topológica na qual neurônios próximos representam características semelhantes nos dados. Cada amostra de entrada é associada ao neurônio vencedor, cujo peso é o mais próximo da amostra no espaço de entrada. Os neurônios vencedores são fundamentais para a interpretação dos resultados finais do SOM, pois indicam a localização dos clusters no mapa SOM, facilitando a análise da distribuição espacial dos dados e a identificação de padrões complexos.

Para combinar o SOM com K-means, primeiro utilizei o SOM para organizar as imagens das células em uma grade 2D. Cada imagem foi mapeada para o seu respectivo neurônio vencedor na grade

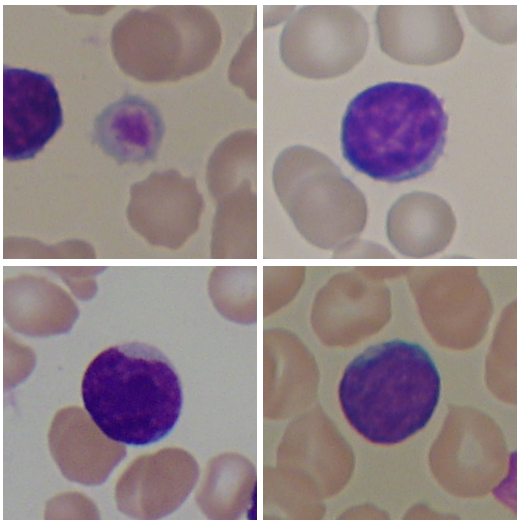

Figure 1. Amostras das Células: na parte superior, células não saudáveis; na parte inferior, células saudáveis.

SOM. Em seguida, apliquei o K-means aos neurônios vencedores do SOM, tratando-os como pontos no espaço de características reduzido obtido após a aplicação do PCA. Essa abordagem permitiu a formação de clusters finais utilizando a informação de proximidade espacial organizada pelo SOM, visando melhorar a precisão na separação entre células saudáveis e não saudáveis.

Para interpretar os resultados finais do SOM, é comum a utilização da Matriz U.

A Matriz U é uma ferramenta visual para interpretar o produto final do SOM. Essa matriz representa a distância média entre os neurônios vizinhos no mapa, fornecendo informações sobre a organização espacial dos dados e a densidade dos clusters. A construção da Matriz U envolve o cálculo das distâncias euclidianas entre os pesos dos neurônios vizinhos e é fundamental para compreender a topologia emergente do mapa.

Cada célula na Matriz U é colorida de acordo com a distância média entre um neurônio e seus vizinhos. Cores mais claras indicam regiões onde os neurônios estão mais distantes uns dos outros, enquanto cores mais escuras indicam regiões onde os neurônios estão mais próximos, formando agrupamentos mais densos.

2.2. K-means

O K-means é uma técnica de clustering que visa particionar um conjunto de dados $\mathbf{X}$ em k grupos distintos, onde k é um número previamente definido. O K-means busca minimizar a função objetivo:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Onde C_i é o conjunto de pontos atribuídos ao cluster i e μ_i é o centroide do i-ésimo cluster.

Inicialmente, são selecionados k centroides aleatórios a partir dos dados $\mathbf{X}$. Em seguida, cada ponto de dados é atribuído ao cluster cujo centroide está mais próximo, com base na distância Euclidiana. Os centroides são então recalculados como a média dos pontos atribuídos a cada cluster. Esse processo de atribuição e atualização dos centroides é repetido iterativamente até que não haja mudanças significativas nos centroides ou até que um critério de parada seja atingido.

Para determinar o número ideal de clusters k, utilizamos o método do cotovelo. Esse método plota a soma dos erros quadráticos dentro dos clusters (Within-Cluster Sum of Squares, WCSS) em função do número de clusters k. O ponto onde a diminuição de WCSS começa a se estabilizar, formando um ”cotovelo”, indica o número ideal de clusters. Esse ponto de inflexão sugere um valor de k onde a adição de mais clusters não resulta em uma redução significativa do WCSS.

Adicionalmente, para confirmar o número ideal de clusters identificado pelo método do cotovelo, utilizamos o coeficiente de silhueta. A silhueta mede a coesão e a

separação dos clusters, com valores que variam de $[-1, 1]$. Um valor de silhueta próximo de 1 indica que os pontos estão bem agrupados dentro dos seus clusters e bem separados dos outros clusters.

Para a análise dos grupos, as medidas de distância intra-cluster e inter-cluster foram utilizadas. A distância intra-cluster mede a coesão dentro dos clusters, onde valores menores indicam maior coesão. A distância inter-cluster vai medir a separação entre os clusters, onde valores maiores indicam uma melhor distinção.

2.3. Análise de Componentes Principais

Análise de Componentes Principais (PCA) visa transformar um conjunto de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas, chamadas componentes principais (CP). Esses CP capturam parte da variabilidade presente nos dados, permitindo a representação dos dados em um espaço de dimensão menor.

Antes de aplicar técnicas como o SOM e o K-means, é comum a utilização do PCA para reduzir a dimensionalidade dos dados e extrair as características mais importantes.

3. Resultados

3.1. SOM+K-means

A seguir, apresentamos a grade resultante do SOM, implementada em uma estrutura de tamanho 20×20, i.e, com 400 neurônios. Utilizamos um comprimento de entrada de 50 após a aplicação do PCA, escolhido para capturar as principais variações nos dados de imagens. O comprimento de entrada refere-se ao número de características ou dimensões dos dados que são fornecidos como entrada para o algoritmo SOM. Os parâmetros $\sigma = 2$ e taxa de aprendizado $\alpha = 0.9$ foram ajustados para controlar o tamanho da vizinhança da grade. Além disso, o algoritmo foi treinado ao longo de 100.000 iterações.

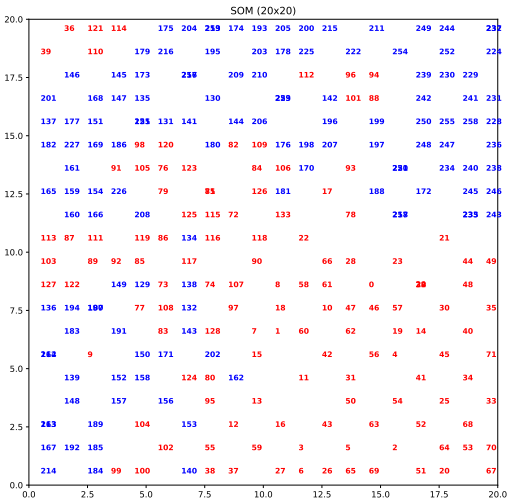

Figure 2. Grade SOM.

O dataset usado neste estudo inclui rótulos verdadeiros, onde as imagens de 1 a 130 correspondem a células não saudáveis, que são marcadas em vermelho, enquanto as imagens de 131 a 260 representam células saudáveis, que são marcadas em azul. O agrupamento das células não saudáveis na parte inferior direita da grade foi realizado de maneira satisfatória. Da mesma forma, na parte superior da grade, observamos um agrupamento homogêneo das células saudáveis, onde há apenas 10 células não saudáveis agrupadas incorretamente.

De maneira geral, o visual que esperamos ver na grade é a reunião de pontos da mesma cor. No entanto, notamos uma sobreposição de cores na parte central da grade do SOM, à esquerda, indicando uma dificuldade em distinguir entre os diferentes tipos de células. Para uma análise visual mais detalhada, apresentamos a Matriz-U.

Na Matriz-U, observamos detalhadamente a separação das células realizada pelo SOM. A parte superior direita apresenta tons mais escuros, indicando áreas com agrupamentos mais densos. Além disso, os pontos mais claros na matriz obtida estão

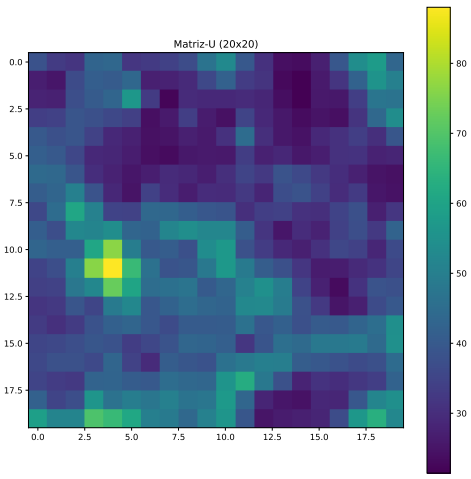

Figure 3. Matriz-U.

localizados à esquerda, onde ocorre uma maior sobreposição entre os clusters de células saudáveis e não saudáveis.

As medidas de desempenho para o SOM+K-means revelam um valor de Intra-cluster de 127.44 e Inter-cluster de 140.51. O Intra-cluster indica a coesão dos agrupamentos formados. Já o Inter-cluster representa uma medida da separação entre diferentes clusters.

3.2. K-means

Inicialmente, exploramos o desempenho do algoritmo K-means sem a prévia organização dos dados pelo SOM. Optamos por determinar o número ideal de clusters, k, utilizando o método do cotovelo. Na Figura 4, o ponto de inflexão sugere que o número ideal de clusters está no intervalo de $[2, 5]$.

Para confirmar a escolha de k, calculamos o coeficiente de silhueta para os valores de k sugeridos pelo método do cotovelo. A análise da silhueta mostrou que o valor ótimo de k é de fato 2, com um índice de silhueta de 0.4535. Os índices de silhueta para os outros valores de k foram 0.20 para

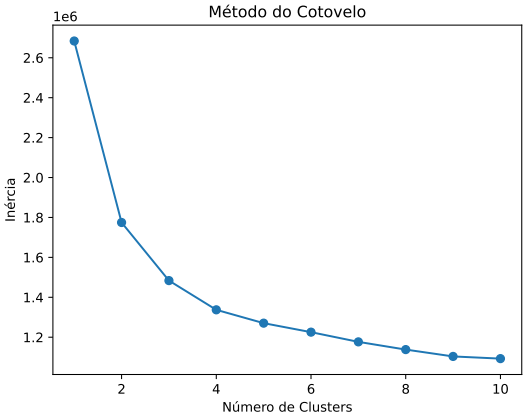

Figure 4. Método do Cotovelo.

$k = 3$, 0.18 para $k = 4$, e 0.19 para $k = 5$.

O valor da silhueta de 0.4535 sugere que os grupos são razoavelmente bem definidos, entretanto, há alguma sobreposição entre eles. O gráfico das duas CP do K-means é mostrado na Figura 5.

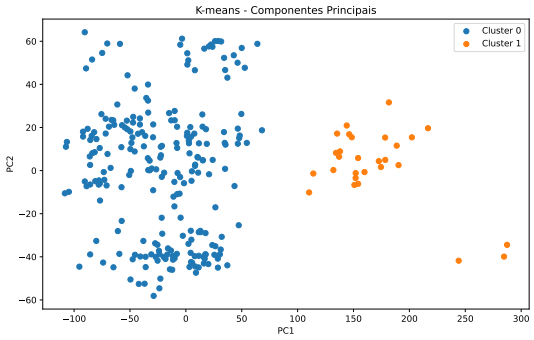

Figure 5. K-means.

Por fim, o K-means apresentou um valor Intra-cluster de 103.96 e Inter-cluster de 213.24.

3.3. Resumo

O SOM+K-means apresentou as seguintes métricas: Intra-cluster de 127.44 e Inter-cluster de 140.51, enquanto o K-means mostrou Intra-cluster de 103.96 e Inter-cluster de 213.24.

A Figura 6 ilustra as diferenças nas métricas entre as abordagens. Embora

o SOM+K-means identifique um maior número de tipos de células, ele apresenta uma sobreposição significativa entre os clusters, o que pode dificultar a distinção clara entre os dois grupos.

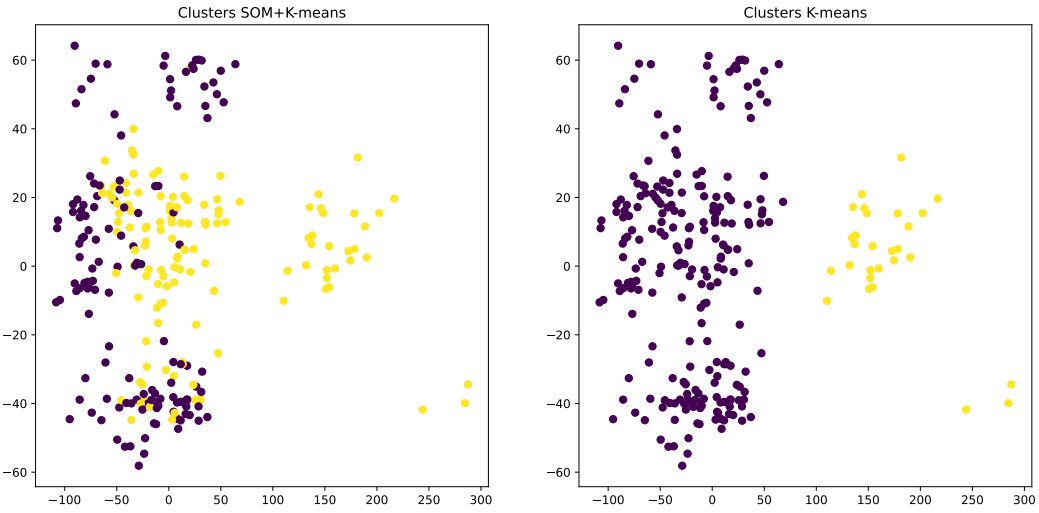

Figure 6. S+K vs K-means.

4. Conclusão

A abordagem SOM+K-means distinguiu melhor os dois tipos de células (Figura 6), apesar de sobrepor significativamente os clusters, como foi indicado pelas métricas. Por outro lado, o K-means (isolado) mostrou-se superior construindo clusters mais coesos e melhor separados entre si. Para aplicações que requerem uma clara distinção entre diferentes grupos de células, o K-means isolado pode ser preferível.

No entanto, há potencial para melhorar a abordagem combinada SOM+K-means. Considerações futuras poderiam explorar o uso de uma rede neural convolucional pré-treinada para extração de características mais discriminativas das imagens. Além disso, um mapa SOM com mais neurônios poderia oferecer uma representação mais detalhada e menos sobreposta dos dois tipos de células, melhorando assim a eficácia do agrupamento.

5. Referências

Izenman, J. A. 2008. Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. Springer.