Paradigma Não-Supervisionado

Ana Raquel Romeu Aguiar   
*Centro de Ciências Tecnológicas*  
*Universidade de Fortaleza*Fortaleza, Brasil  
<https://orcid.org/0000-0003-0821-3302>   
Matrícula: 2517191  
Sidney Ferreira Baptista Junior  
*Centro de Ciências Tecnológicas*  
*Universidade de Fortaleza*  
*Fortaleza, Brasil*[*sidneyjunr@edu.unifor.br*](mailto:sidneyjunr@edu.unifor.br) *Matrícula: 2425146*

##### *X. Aprendizado Não-Supervisionado*

O conjunto de dados utilizado para o aprendizado não-supervisionado é de 640 fotos de diversas pessoas em diversas poses, tudo dividido em múltiplas pastas. As fotos têm as mesmas dimensões: 128 x 120, ou seja, ao multiplicar essas dimensões, o conjunto é composto de 15.360 variáveis independentes. Portanto, o uso do aprendizado supervisionado se torna inviável pela quantidade elevada de variáveis, o que torna o processo demorado e caro [1].

Outro motivo para se escolher a metodologia do aprendizado não-supervisionado neste caso é a falta de informação relacionada aos dados [2], já que não se sabe muito além de suas dimensões e da sua natureza. São fotos de quê? Elas têm algum elemento em comum? É possível dividir todas essas fotos em grupos menores? O aprendizado não-supervisionado ajudará a responder tais perguntas.

Como a quantidade de dados (mais de 15 mil) ainda é extensa, é preciso aplicar uma Redução de Dimensionalidade, ou seja, um processo para minimizar a quantidade de variáveis a fim de facilitar a análise de dados, omitir possíveis redundâncias nos dados e aumentar a eficiência do aprendizado [3].

Três técnicas foram utilizadas para fazer essa redução: t-SNE (*t-Distributed Stochastic Neighbor Embedding*), PCA (Análise de componentes principais) e UMAP (*Uniform Manifold Approximation and Projection*). Os gráficos com os resultados de cada uma das técnicas serão mostrados neste artigo. Em seguida, os resultados serão avaliados para eleger a técnica escolhida para a próxima fase, a de agrupamento.

Para o t-SNE, foi criado um gráfico (Fig 1) de dois componentes (tsne = TSNE(n\_components=2)), visto que a visualização dos dados e do agrupamento inicial é mais apropriada para um artigo que um gráfico tridimensional (n\_components=3).

1. Gráfico do t-SNE, componentes = 2

O resultado exibido mostra um conjunto de vários subgrupos, porém a quantidade e distribuição ainda parece muito complicada para prosseguir com o agrupamento.

Em seguida, o gráfico do PCA (Fig 2) indica o ponto onde a amostra pode trazer informação redundante.

1. Gráfico do PCA padrão (100%)

É possível perceber que a curvatura do gráfico muda em um ponto em que x está próximo de 100, o que se infere que as 500 amostras restantes podem trazer menos informação relevante sobre os dados.

É a partir desse pressuposto que será feita a redução de dimensionalidade com o PCA. Três porcentagens de variância foram escolhidas para a análise: 90%, 80% e 75%. Para isso, é necessário avaliar, no gráfico, qual o valor de x quando o valor de y é 0,9, 0,8 e 0,75, respectivamente.

Quando y= 0,9, x= 71,77. Ao arredondar esse valor, x = 72. Esse valor será usado no campo de número de componentes (pca = PCA(n\_components=72)). Percebe-se imediatamente a redução da amostra de 640 a 72. Esse valor indica que esses 72 componentes contêm 90% da informação do conjunto. A Fig. 3 mostra o gráfico com essa variância.

1. Gráfico do PCA com variância de 90%

Esse processo se repete para as outras duas variâncias. Quando y=0,8, x=29,62, que é arredondado para 30. Neste caso, o número de componentes é mudado para 30 (n\_components=30) e gera o gráfico da Fig D.

1. Gráfico do PCA com variância de 80%

Finalizando a parte de PCA, a variância de 75%. Quando y=0,75, x= 21,68, arrendodado para 22. Ao configurar o n=componente para 22, é mostrado o gráfico da Fig E.

1. Gráfico do PCA com variância de 75%

A seguir, será feita a redução pelo UMAP, que funciona de forma semelhante ao t-SNE, mas mais eficiente e com mais preservação da estrutura global da amostra [4]. Para avaliação inicial, o número de componentes foi 2 (umap\_var = umap.UMAP(n\_components=2)) e gerado o gráfico a seguir (Fig F).

1. Gráfico do UMAP com componentes = 2

Ao comparar ao gráfico com o t-SNE (Fig A), percebe-se que as amostras estão bem menos espalhadas, o que já facilita alguns agrupamentos.

A seguir, o teste será feito com as dimensões 3, 15, 55 e 101. Esses números serão incluídos no código no campo de número de componentes, substituindo o 2.

Para n\_components=3, o gráfico gerado é o da Fig G.

1. Gráfico do UMAP com componentes = 3

As amostras, agora, estão divididas em três grandes grupos, cada um com um número diversos de subgrupos. Comparado aos gráficos anteriores, a divisão é mais evidente.

A seguir, o gráfico da Fig H mostra o UMAP com componentes = 15.

1. Gráfico do UMAP com componentes = 15

Agora, há dois grupos pequenos isolados e um grande grupo com as amostras restantes. Pode ser interessante analisar os isolados, mas o grupo maior ainda pode ser problemático.

A Fig I mostra como seria a divisão com componentes = 55.

1. Gráfico do UMAP com componentes = 55

A estrutura do gráfico é parecida com a anterior (Fig H) (dois grupos pequenos e um maior), mas a direção está contrária.

Finalmente, o gráfico com componentes = 101 encontra-se a seguir (Fig J).

1. Gráfico do UMAP com componentes = 101

Não há grandes diferenças entre as figuras I e J, além de uma separação um pouco maior entre alguns grupos pequenos.

Após análise de todos os gráficos, a técnica UMAP com 3 componentes foi escolhida por propor o melhor agrupamento para o caso deste estudo.

1. *Agrupamento (Clustering)*

Com a etapa de redução de dimensionalidade concluída, a próxima fase consiste na aplicação de métodos de aprendizado não-supervisionado para identificar padrões e grupos presentes nos dados. Neste estudo, foram utilizados os algoritmos K-means e K-medoids, ambos adequados para conjuntos de dados sem rótulos e com características de alta dimensionalidade.

1. *2.1. Aplicação do K-means*

O K-means é um algoritmo de clusterização baseado em centróides, que busca minimizar a soma das distâncias quadráticas entre os pontos e o centróide do seu respectivo cluster. Foram testados diferentes valores de k (número de clusters), variando de 2 a 6, com o objetivo de avaliar qual configuração proporciona melhor separação e coesão dos grupos.

Para cada k, foram calculadas as métricas:

* Índice de Dunn, que avalia a relação entre a distância mínima entre clusters e a dispersão máxima intra-cluster;
* Silhouette Score, que mede o quão próximo cada ponto está de seu cluster em relação aos demais clusters.

Os resultados obtidos indicam que o valor k=3 apresentou melhor equilíbrio entre separação e coesão dos grupos, conforme mostrado nas Figuras K e L, que ilustram os valores de Silhouette e Dunn para cada valor de k.

1. *Aplicação do K-medoids*

O K-medoids é similar ao K-means, porém utiliza medoides (pontos reais do conjunto de dados) como representações centrais, tornando-o mais robusto a outliers. A implementação também foi realizada para k= 2 a 6, e as mesmas métricas (Dunn e Silhouette) foram calculadas para cada configuração.

Os resultados obtidos pelo K-medoids foram consistentes com os do K-means, com destaque para sua robustez diante de possíveis pontos isolados no conjunto de dados.

1. *Comparação Visual e Métricas*

A Figura M apresenta a comparação visual entre os clusters identificados pelo K-means e pelo K-medoids para k=3. Observa-se que ambos os algoritmos produziram agrupamentos semelhantes, indicando que os clusters no espaço reduzido pelo UMAP 55D são bem definidos e estáveis.

Além disso, as métricas confirmam a qualidade dos agrupamentos:

| Algoritmo | Silhouette Score | Índice de Dunn | ARI |
| --- | --- | --- | --- |
| K-means | 0.673 | 0.635 | 1.0 |
| K-medoids | 0.673 | 0.635 | 1.0 |

O ARI (Adjusted Rand Index) de 1.0 evidencia que ambos os métodos encontraram a mesma partição dos dados, corroborando a confiabilidade dos clusters obtidos.

1. 2.4. Discussão dos Resultados

Os resultados obtidos demonstram que a escolha da redução de dimensionalidade foi adequada, permitindo que os algoritmos de clusterização identifiquem grupos de forma eficiente, mesmo em um espaço originalmente com mais de 15 mil variáveis.

O K-means apresentou vantagem em termos de desempenho computacional, sendo mais rápido na convergência, enquanto o K-medoids destacou-se pela robustez a outliers, embora a diferença prática não tenha se manifestado neste conjunto de dados específico.

A escolha de k=3 é justificada pelo equilíbrio entre separação e coesão dos clusters, refletido tanto nas métricas quantitativas quanto na análise visual dos agrupamentos. Dessa forma, a etapa de clusterização fornece uma base sólida para futuras análises, podendo ser utilizada para classificação posterior, reconhecimento de padrões ou identificação de subgrupos relevantes.

1. *3. Conclusão Geral*

O presente estudo abordou a aplicação de aprendizado não-supervisionado em um conjunto de dados composto por 640 imagens de diferentes pessoas, cada uma com 15.360 variáveis independentes. Devido à elevada dimensionalidade e à ausência de rótulos, foi necessário aplicar uma etapa de redução de dimensionalidade, seguida da clusterização para identificar padrões e agrupamentos nos dados.

1. 3.1. Redução de Dimensionalidade

Foram testadas três técnicas principais: t-SNE, PCA e UMAP. Os resultados indicaram que:

* O t-SNE apresentou uma boa visualização inicial em duas dimensões, mas não preservou de forma robusta a estrutura global do conjunto de dados.
* O PCA, ao reduzir a dimensionalidade para 90%, 80% e 75% da variância, mostrou-se eficaz na eliminação de redundâncias, embora as visualizações não fossem suficientes para identificar claramente grupos separados.
* O UMAP, em especial na configuração com 55 componentes, apresentou a melhor preservação da estrutura local e global, facilitando a visualização de agrupamentos consistentes.

Dessa forma, a escolha da redução de dimensionalidade via UMAP (55D) foi justificada pela clareza visual dos clusters e pela eficiência na etapa subsequente de clusterização.

1. 3.2. Clusterização

Para a identificação de grupos, foram aplicados os algoritmos K-means e K-medoids. Ambos apresentaram resultados similares, com ARI = 1.0, indicando que a mesma partição de dados foi encontrada independentemente do método.

As métricas de validação (Silhouette e Dunn) confirmaram que k = 3 é o número de clusters mais adequado, proporcionando um bom equilíbrio entre separação e coesão dos grupos.

* O K-means destacou-se por sua rapidez computacional.
* O K-medoids ofereceu maior robustez contra outliers, embora não tenha havido diferenças significativas neste conjunto específico.

1. 3.3. Considerações Finais

O estudo demonstra que, mesmo em conjuntos de dados de alta dimensionalidade e complexidade, é possível identificar padrões significativos utilizando técnicas de redução de dimensionalidade combinadas com métodos de clusterização. A análise visual aliada às métricas quantitativas garantiu a confiabilidade dos agrupamentos.

Este processo não apenas facilita a compreensão da estrutura interna dos dados, mas também cria uma base para futuros estudos de classificação, reconhecimento de padrões ou análises exploratórias, tornando o aprendizado não-supervisionado uma ferramenta poderosa em contextos com dados pouco conhecidos ou não rotulados.

##### References

1. LENZ, Maikon Lucian et al. **Fundamentos de aprendizagem de máquina**. Porto Alegre: Sagah, 2020. 1 recurso online. (Inteligência artificial). ISBN 9786556900902. P. 36. Disponível em:. Acesso em: 31 ago. 2025.
2. LENZ, Maikon Lucian et al. **Fundamentos de aprendizagem de máquina**. Porto Alegre: Sagah, 2020. 1 recurso online. (Inteligência artificial). ISBN 9786556900902. P. 46. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9786556900902>. Acesso em: 31 ago. 2025.
3. VIANNA, Daniel Ribeiro Alves Barboza; MAGALHÃES, Claudio Freitas de; MINEIRO, Érico Franco. Classificação não supervisionada de imagens em processos de design computacional generativo. **Gestão & Tecnologia de Projetos**, São Carlos, v. 19, n. 3, p. 27–47, (p.32-33) 2024. DOI: 10.11606/gtp.v19i3.227541. Disponível em: <https://revistas.usp.br/gestaodeprojetos/article/view/227541>. Acesso em: 31 ago. 2025.
4. MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. (p. 2) Disponível em: <https://arxiv.org/abs/1802.03426>. Acesso em: 01 set. 2025.