FLÁIRA HANNY BOMFIM DOS SANTOS
LIS LOUREIRO SOUSA

RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DE CLASSIFICAÇÃO COM REDES CONVOLUCIONAIS E O DATASET CUFS

VITÓRIA DA CONQUISTA
02/12/2024

1. Resumo

O presente estudo investiga padrões em dados de reconhecimento de atividades humanas por meio do algoritmo de agrupamento K-means. Para isso, os dados foram submetidos a um processo de normalização e redução de dimensionalidade utilizando a Análise de Componentes Principais (PCA), de forma a facilitar a formação de agrupamentos coesos. O número ideal de clusters foi determinado por dois métodos distintos: o Método do Cotovelo e a Métrica de Silhouette. Os resultados obtidos demonstraram uma separação satisfatória entre os clusters, avaliada por meio de métricas quantitativas e visualizações gráficas. Apesar da eficácia do modelo, foram observadas limitações, como a dependência do PCA para interpretação das variáveis originais e a subjetividade na definição do número de clusters, o que abre caminhos para aprimoramentos futuros.

2. Introdução

O reconhecimento de atividades humanas com base em dados de sensores é um desafio relevante em diversas áreas, como saúde, esportes e segurança. No entanto, a complexidade inerente a esses dados, geralmente de alta dimensionalidade e com presença de ruído, exige o uso de técnicas avançadas de aprendizado de máquina. O algoritmo K-means se destaca como uma abordagem eficaz para identificação de padrões, sendo amplamente utilizado em aplicações não supervisionadas devido à sua simplicidade e eficiência. Neste contexto, a aplicação da Análise de Componentes Principais (PCA) justifica-se pela necessidade de reduzir a dimensionalidade dos dados, minimizando o impacto do ruído e facilitando a formação de agrupamentos compactos e bem definidos. Assim, este trabalho explora a performance do K-means aplicado a dados de reconhecimento de atividades humanas, avaliando os resultados em termos de coesão e separação dos clusters.

Metodologia

O estudo iniciou-se com o pré-processamento dos dados, composto por 561 variáveis, as quais foram normalizadas utilizando a técnica *StandardScaler* para garantir a uniformidade na escala das características. Posteriormente, os dados foram submetidos à Análise de Componentes Principais (PCA), que reduziu a dimensionalidade para 50 componentes principais, preservando uma proporção significativa da variância original.

Para determinar o número ideal de clusters, foram adotados dois métodos complementares. O primeiro, conhecido como Método do Cotovelo, avaliou a inércia dos agrupamentos para identificar um ponto de inflexão, sugerindo o valor mais adequado para KK. Já o segundo método, baseado na Métrica de Silhouette, considerou a maximização da separação média entre os clusters, fornecendo uma alternativa quantitativa para a escolha do número de agrupamentos.

O algoritmo K-means foi implementado utilizando a variante *MiniBatchKMeans*, que é mais eficiente computacionalmente para grandes conjuntos de dados. Além disso, múltiplas execuções foram realizadas para avaliar a consistência dos agrupamentos formados. Por fim, gráficos e análises descritivas foram empregados para validar e interpretar os resultados obtidos.

3. Resultados

A análise indicou um único valor para o número ideal de clusters. Pelo Método do Cotovelo, identificou-se um ponto de inflexão na curva de inércia, sugerindo um valor apropriado para KK. Por outro lado, a Métrica de Silhouette apontou outro valor para KK, baseado na separação média entre clusters.

Os resultados mostraram que os clusters formados apresentaram uma coesão satisfatória, evidenciada por métricas quantitativas e visualizações em duas dimensões, realizadas após a redução de dimensionalidade adicional com PCA. Além disso, a alta consistência observada em execuções repetidas validou a robustez do modelo.

Análises descritivas indicaram que as atividades humanas não se distribuem uniformemente entre os clusters, sugerindo a presença de padrões específicos nos dados. Apesar disso, a dependência da redução de dimensionalidade limitou a interpretação direta das variáveis originais no contexto dos agrupamentos.

4. Discussão

Os resultados obtidos demonstram que o K-means, combinado com PCA, é uma ferramenta eficaz para a identificação de padrões em dados de alta dimensionalidade. A coesão e separação dos clusters foram satisfatórias, como evidenciado pelos gráficos e métricas

utilizadas. Contudo, a aplicação de PCA pode ter impactado negativamente a interpretabilidade das características originais, o que representa uma limitação metodológica.

A subjetividade envolvida no Método do Cotovelo é outro ponto de atenção, uma vez que a escolha do número de clusters depende de interpretação visual do gráfico de inércia. Além disso, a presença de ruído nos dados sensoriais pode ter influenciado a qualidade dos agrupamentos formados. Apesar disso, a adoção de *MiniBatchKMeans* foi uma escolha acertada, contribuindo para a eficiência computacional do processo.

De forma geral, as escolhas metodológicas mostraram-se adequadas para os objetivos do estudo, mas ressaltam a necessidade de abordagens complementares para validar e enriquecer as análises realizadas.

5. Análise dos Resultados Obtidos com os Gráficos

O primeiro gráfico apresenta a variação do índice *Silhouette* em função do número de clusters (K), o que permite avaliar a coesão e a separação dos grupos gerados pelo algoritmo K-means. Observa-se que o *Silhouette Score* atinge seu valor mais elevado para K=2K=2, sugerindo que esta é a quantidade ideal de clusters para o conjunto de dados analisado. Esse comportamento reflete a capacidade do modelo em formar grupos bem definidos, com maior proximidade interna entre os pontos de cada cluster e uma separação mais clara entre os diferentes clusters. À medida que o número de clusters aumenta, o índice diminui progressivamente, indicando uma piora na qualidade da segmentação.

O segundo gráfico visualiza os clusters formados para K=2K=2 com a redução dimensional realizada por Análise de Componentes Principais (PCA). Os dados foram projetados em um espaço bidimensional, onde cada cluster é representado por cores distintas. O gráfico evidencia dois grupos bem separados, com padrões claros de distribuição espacial que corroboram os resultados obtidos pelo índice *Silhouette*. A diferença entre os clusters sugere que o modelo conseguiu capturar adequadamente a estrutura dos dados, separando-os de forma lógica e coerente.

Portanto, a combinação dos gráficos reforça a escolha de K=2K=2 como o ponto de maior eficiência do algoritmo K-means aplicado a este problema.

Por fim, O gráfico apresentado ilustra os resultados de agrupamento (clustering) realizados sobre um conjunto de dados reduzido para duas dimensões por meio da técnica de Análise de Componentes Principais (PCA). Os clusters foram obtidos utilizando o método do cotovelo para determinar o número ideal de agrupamentos, definido como K=2K = 2K=2. A escolha deste KKK foi baseada no ponto de inflexão observado no gráfico da inércia ao longo dos valores de KKK.

Resultados Observados

1. **Clusters bem definidos**: No gráfico, é possível identificar dois agrupamentos principais, destacados por cores distintas (roxo e amarelo). Esses clusters indicam que os dados foram separáveis em dois grupos com características semelhantes dentro de cada grupo.
2. **Separação com PCA**: A redução dimensional para duas componentes principais permitiu visualizar claramente os clusters. A Componente Principal 1 parece ser a variável com maior variância explicada, sendo responsável pela maior separação entre os grupos.
3. **Método do Cotovelo**: A análise do cotovelo indicou K=2K = 2K=2 como o número ideal de clusters. Isso é coerente com o gráfico, que mostra uma clara separação entre os dois grupos principais.
4. **Interpretação dos dados**: Os clusters podem estar relacionados a padrões subjacentes nas características dos dados originais. Com base no código analisado, esses dados vêm do conjunto UCI HAR Dataset, utilizado para reconhecimento de atividades humanas.

Impacto dos Resultados

- **Performance computacional**: A utilização de MiniBatchKMeans em vez de KMeans padrão permitiu um desempenho otimizado para grandes volumes de dados.
- **Visualização clara**: A representação em 2D com PCA facilita a comunicação dos resultados e permite uma análise qualitativa da separação dos clusters.

Possíveis Limitações

- **Simplicidade do modelo**: Apesar da boa separação visual, K=2K = 2K=2 pode ser simplista para problemas mais complexos. Análises adicionais com outras métricas, como o Silhouette Score, poderiam complementar essa escolha.
- **Redução de dimensionalidade**: Embora o PCA preserve a variância, informações críticas para interpretação podem ser perdidas durante a redução.

Conclusão

O agrupamento realizado revelou dois grupos bem definidos no espaço dimensional reduzido. Esse processo evidencia a eficiência do método de cotovelo e do PCA para analisar e visualizar grandes conjuntos de dados. Aplicações práticas podem incluir classificação de atividades, detecção de padrões anômalos ou agrupamento para análise exploratória.

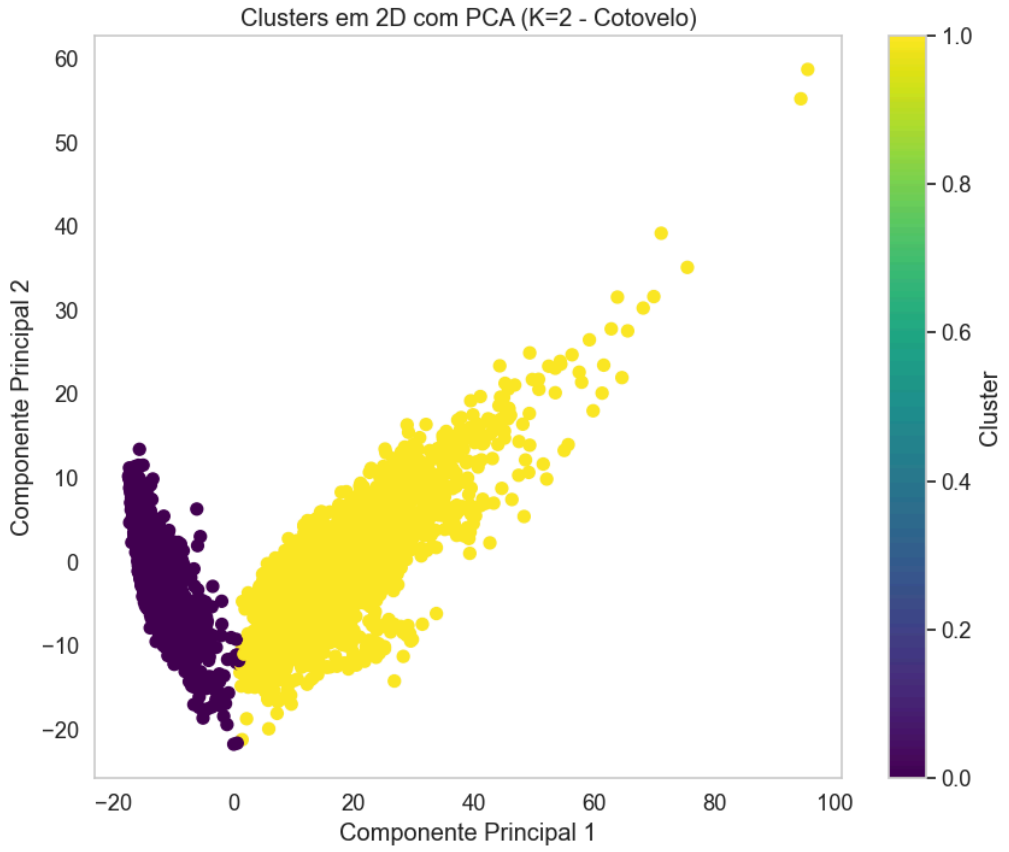

Gráfico 01

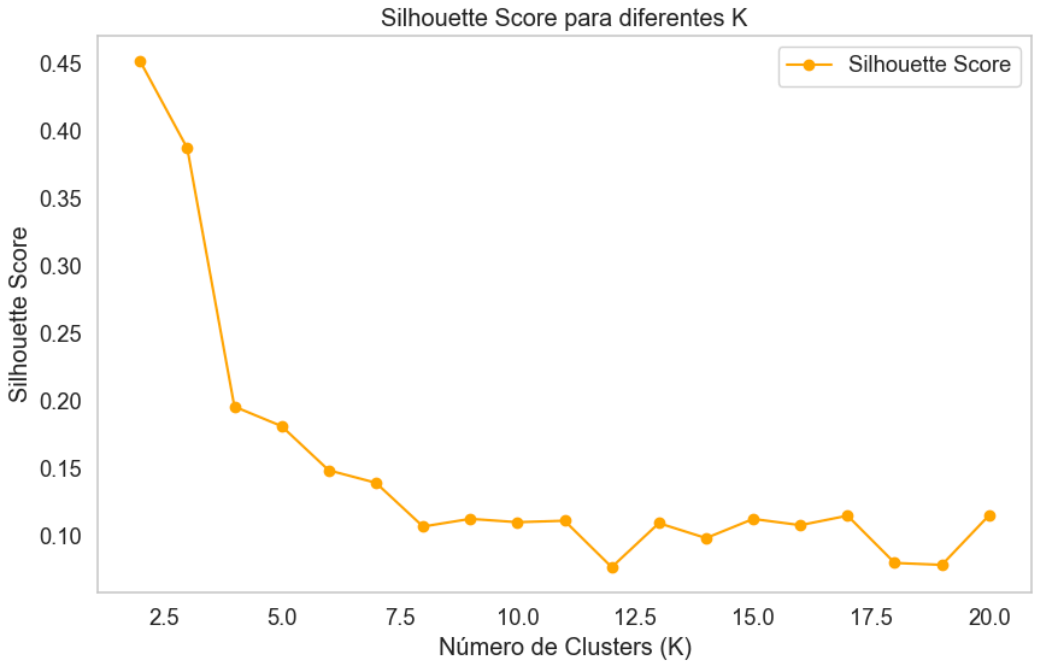

Gráfico 02

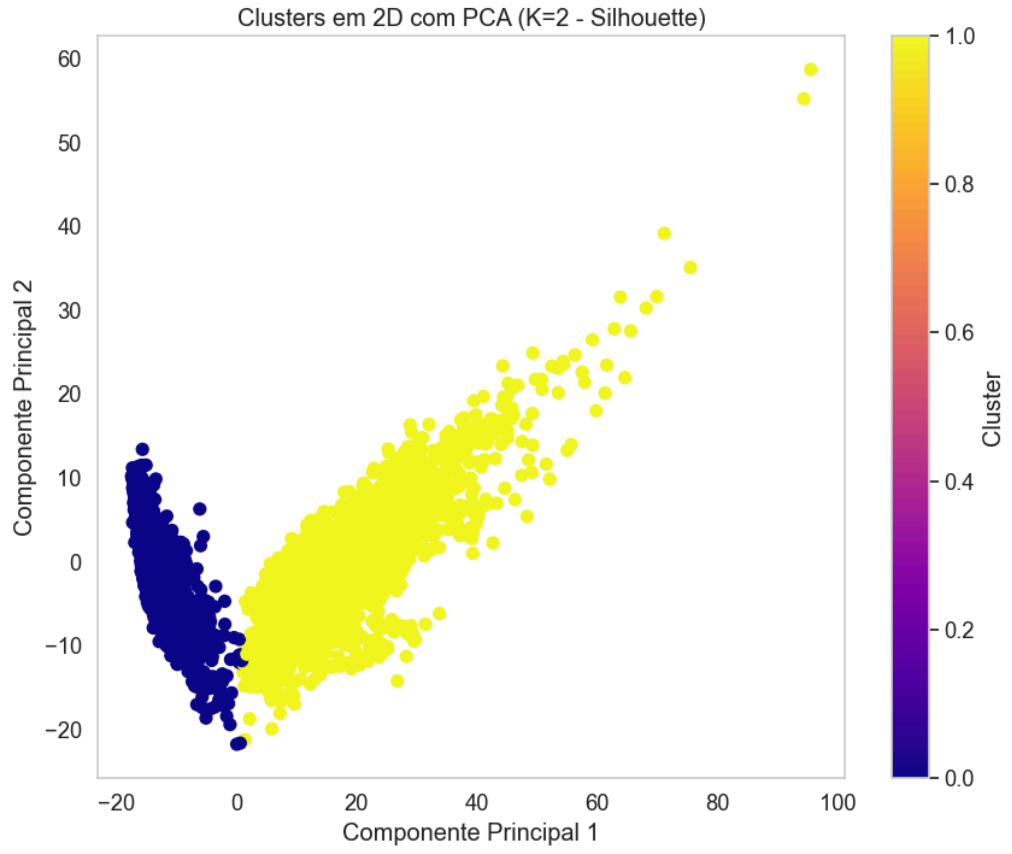

Gráfico 03

6. Conclusão e Trabalhos Futuros

Este estudo demonstrou a viabilidade do uso de K-means em conjunto com PCA para explorar padrões em dados de reconhecimento de atividades humanas. Os resultados obtidos indicam que é possível formar clusters coesos e interpretáveis, ainda que algumas limitações tenham sido observadas, como a perda de interpretabilidade das variáveis originais e a subjetividade em algumas decisões metodológicas.

Para trabalhos futuros, sugere-se explorar algoritmos alternativos, como o DBSCAN, que é mais robusto para clusters de formas não esféricas. Além disso, incorporar informações temporais e contextuais poderia enriquecer a análise, possibilitando uma compreensão mais profunda dos padrões identificados. Por fim, estratégias híbridas que combinem aprendizado supervisionado e não supervisionado podem oferecer validação adicional e aprimorar os resultados obtidos.

7. Referencias

K-means clustering. Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/K-means_clustering>. Acesso em: 3 dez. 2024.

K-means. Wikipedia. Disponível em: <https://pt.wikipedia.org/wiki/K-means>. Acesso em: 3 dez. 2024.

Dendrogram. Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/Dendrogram>. Acesso em: 3 dez. 2024.

Cluster analysis. Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/Cluster_analysis>. Acesso em: 3 dez. 2024.