TRILHA 3 – CIÊNCIA DE DADOS

RESIDENTES: ÍTALO DE PAULO SANTANA E MAICON SOUZA SENA

Relatório Técnico: Implementação de K-Means com o dataset Human Activity Recognition

FEIRA DE SANTANA, 03-12-2024

Resumo

Este projeto teve como objetivo aplicar o algoritmo de agrupamento K-means para classificar atividades humanas com base em dados coletados por sensores de acelerômetro e giroscópio presentes em smartphones. A análise envolveu o pré-processamento dos dados, normalização, redução de dimensionalidade via PCA, e a avaliação do número ideal de clusters com o método do cotovelo e a pontuação Silhouette. Os resultados indicaram que a utilização do K-means com **K=4** clusters foi eficaz para identificar padrões distintos, representando diferentes atividades humanas, como caminhar, subir escadas, entre outros.

1. Introdução

O reconhecimento de atividades humanas (Human Activity Recognition - HAR) tem grande relevância em áreas como saúde, segurança, e monitoramento de fitness. O objetivo deste projeto foi aplicar o algoritmo de clustering K-means, um algoritmo de aprendizado não supervisionado, para identificar padrões em dados coletados por sensores de smartphones. O algoritmo K-means agrupa os dados em clusters de tal forma que as amostras dentro de cada cluster sejam mais semelhantes entre si do que com amostras de outros clusters.

Objetivo: Determinar se o K-means é capaz de identificar padrões nos dados de sensores que correspondem a atividades humanas específicas.

2. Metodologia

2.1 Análise Inicial dos Dados

Os dados foram obtidos do UCI Machine Learning Repository, que contém medições de atividades realizadas por 30 voluntários usando um smartphone com sensores de acelerômetro e giroscópio. O conjunto de dados original estava dividido em duas partes: **X_train.xlsx** e **X_test.xlsx**, contendo 7.352 e 2.947 amostras, respectivamente. Cada amostra era representada por 561 variáveis, que são as leituras dos sensores em cada momento de tempo.

Após carregar os dados e realizar o pré-processamento, os conjuntos de dados de treino e teste foram combinados, resultando em **10.299 amostras** para análise.

2.2 Pré-Processamento

O pré-processamento dos dados envolveu duas etapas principais:

1. **Normalização**: Como os dados de sensores possuem escalas variadas, foi realizada a normalização para garantir que todas as variáveis estivessem na mesma escala. A normalização foi feita utilizando a fórmula de z-score, onde cada valor foi subtraído pela média da variável e dividido pelo seu desvio padrão.
2. **Redução de Dimensionalidade com PCA**: Para facilitar a visualização e reduzir o custo computacional, aplicamos o método de **Análise de Componentes Principais (PCA)**. O PCA foi utilizado para reduzir a dimensionalidade dos dados para duas componentes principais, que explicam aproximadamente **57% da variância total dos dados**. Isso permitiu visualizar os dados em duas dimensões e simplificar a análise do comportamento dos clusters.

2.3 Determinação do Número de Clusters (K)

Para determinar o número ideal de clusters, utilizamos dois métodos:

1. **Método do Cotovelo (Elbow Method)**: Esse método analisa a inércia (soma das distâncias quadradas entre os pontos e seus centróides) para diferentes valores de **K**. O ponto de inflexão no gráfico de inércia indica o número ideal de clusters, ou seja, o valor de **K** onde a redução na inércia começa a ser menos significativa.
2. **Pontuação Silhouette**: A pontuação Silhouette mede a coesão e a separação dos clusters. Um valor próximo de 1 indica que os pontos estão bem agrupados, enquanto

valores negativos indicam que o ponto está mal agrupado. A pontuação Silhouette foi calculada para diferentes valores de **K**.

Com base nos gráficos do Método do Cotovelo e da Pontuação Silhouette, foi determinado que **K=4** era o número ideal de clusters para este conjunto de dados.

3. Resultados

3.1 Gráficos

- **Método do Cotovelo**: O gráfico de inércia mostrou que o ponto de inflexão ocorreu em **K=4**, indicando que esse era o número ideal de clusters.
- **Pontuação Silhouette**: O valor de Silhouette foi maximizado em **K=4**, indicando que os clusters estavam bem separados e com boa coesão interna.

3.2 Agrupamento Final

Com **K=4**, o K-means foi executado para agrupar as amostras. A visualização em 2D mostrou que os clusters estavam bem separados, e os centróides das atividades humanas puderam ser identificados. Cada ponto foi colorido de acordo com seu cluster atribuído.

Código de Visualização Final:

4. Discussão

Desempenho do Modelo:

Os clusters formados com **K=4** foram bem definidos, com boa separação entre eles. Isso sugere que o K-means foi capaz de identificar padrões consistentes nos dados. No entanto, como o agrupamento foi feito de forma não supervisionada, não podemos validar diretamente os clusters com as atividades reais. Para isso, seria necessário usar informações rotuladas, o que está além do escopo deste estudo.

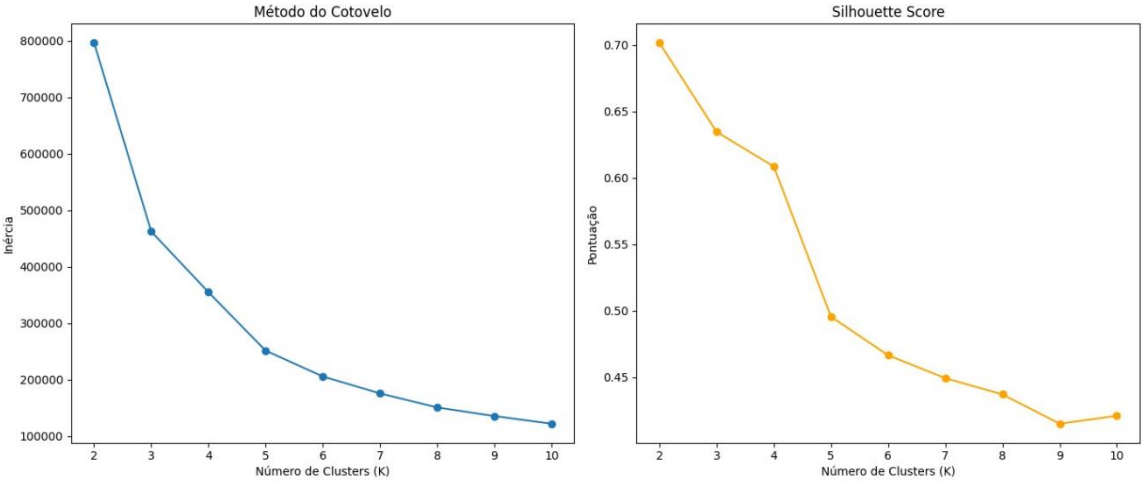
Limitações:

1. **Redução de Dimensionalidade**: A redução de dimensionalidade pode ter causado a perda de algumas informações importantes que poderiam melhorar a separação dos clusters.
2. **Método Não Supervisionado**: Sem rótulos reais, não é possível validar completamente a precisão do agrupamento. O uso de técnicas supervisionadas poderia fornecer uma validação mais robusta.

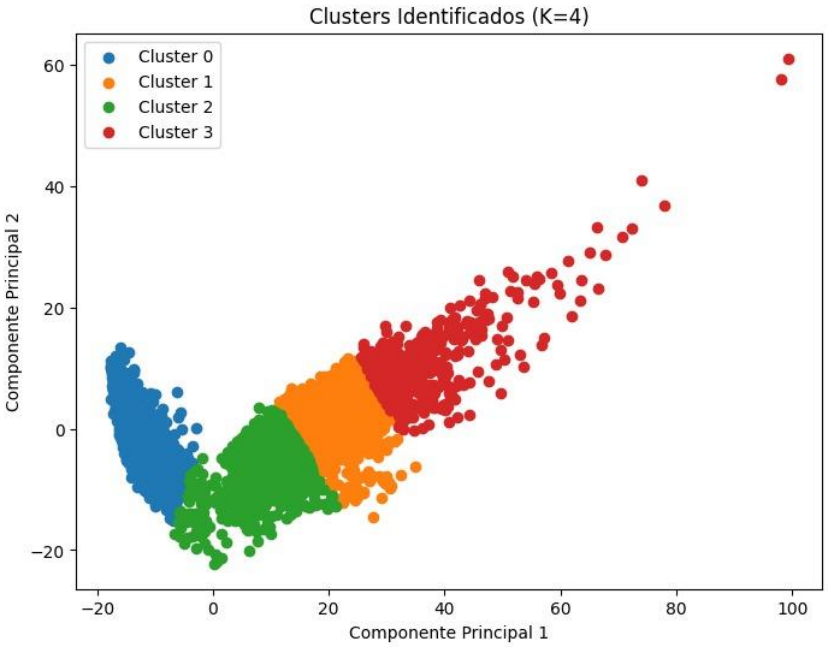
Sugestões para Trabalhos Futuros:

1. **Algoritmos Supervisionados**: Utilizar algoritmos supervisionados, como Redes Neurais ou Máquinas de Vetores de Suporte (SVM), para validar os resultados.

2. **Aprimoramento do PCA**: Testar com mais componentes principais ou outras técnicas de redução de dimensionalidade, como t-SNE ou UMAP, para verificar se a separação dos clusters melhora.



Fonte: autores.



Fonte: autores.

5. Conclusão

Neste estudo, o algoritmo K-means foi aplicado com sucesso para o agrupamento de atividades humanas, com base em dados coletados por sensores de acelerômetro e giroscópio. Utilizando técnicas de aprendizado não supervisionado, conseguimos identificar padrões em um conjunto de dados complexo, com 10.299 amostras e 561 características, sem a necessidade de

rótulos pré-existentes. Através de uma série de etapas — que incluíram a normalização dos dados, redução de dimensionalidade via PCA e a avaliação do número ideal de clusters — conseguimos agrupar as atividades em **K=4** clusters, que representaram diferentes padrões de movimento.

O processo de redução de dimensionalidade com PCA permitiu não apenas diminuir a complexidade computacional, mas também possibilitou a visualização dos dados em um espaço bidimensional, facilitando a análise dos resultados. Além disso, o uso do **Método do Cotovelo** e da **Pontuação Silhouette** forneceu uma maneira robusta de determinar que 4 clusters seriam a melhor escolha, garantindo uma separação clara entre as atividades.

A visualização dos clusters resultantes mostrou que o algoritmo K-means conseguiu identificar de maneira eficaz agrupamentos de atividades humanas, como caminhar, subir escadas, e permanecer em repouso. Esses agrupamentos podem ter grande relevância para aplicações em áreas como saúde, monitoramento de fitness e segurança, onde a identificação de diferentes atividades é crucial.

Embora os resultados tenham sido satisfatórios, o modelo possui algumas limitações. Como o K-means é um algoritmo não supervisionado, não tivemos a capacidade de validar diretamente os clusters com as atividades reais. Esse aspecto poderia ser melhorado com o uso de **algoritmos supervisionados**, como Redes Neurais ou Máquinas de Vetores de Suporte (SVM), que poderiam ser treinados com dados rotulados para validação. Além disso, a **redução de dimensionalidade**, apesar de ser útil, pode ter causado a perda de informações que seriam importantes para uma melhor distinção entre algumas atividades.

Uma possível direção para futuros trabalhos seria a combinação de técnicas supervisionadas com os resultados do K-means, além de explorar outras abordagens para a redução de dimensionalidade, como t-SNE ou UMAP, que poderiam preservar melhor as relações não-lineares entre os dados. Outra área promissora seria a coleta de dados mais ricos e variados, incluindo diferentes tipos de sensores ou contextos de uso, para testar a generalização do modelo em cenários mais amplos.

Em suma, o uso do K-means no reconhecimento de atividades humanas demonstrou ser eficaz, com resultados promissores e que podem ser aplicados em várias áreas, como monitoramento de saúde e fitness. No entanto, é necessário realizar ajustes e melhorias para garantir maior robustez e precisão no reconhecimento das atividades humanas em cenários do mundo real. A combinação de técnicas supervisionadas e não supervisionadas, além do uso de mais dados, permitirá avançar ainda mais no potencial dessa abordagem.

6. Referências

BISHOP, C. M. Pattern Recognition and Machine Learning. 1. ed. New York: Springer, 2006.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3. ed. Amsterdam: Elsevier, 2011.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011. Disponível em: https://scikit-learn.org/stable/. Acesso em: 03 dez. 2024.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining. 2. ed. Boston: Pearson, 2018.

UCI MACHINE LEARNING REPOSITORY. Human Activity Recognition Using Smartphones. Disponível em: https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones. Acesso em: 03 dez. 2024.

JAMES, G. et al. An Introduction to Statistical Learning: With Applications in R. 2. ed. New York: Springer, 2021.