RESIDÊNCIA EM SOFTWARE

BAHIA + TECNOLOGIA + EMPREENDEDORISMO

Projeto: Implementação e Análise do Algoritmo de K-means com o Dataset Human Activity Recognition

Luís Felipe Alves de Moura

01 de dezembro de 2024

Objetivo:

O objetivo deste trabalho foi aplicar o algoritmo de agrupamento **K-means** para reconhecer e agrupar diferentes atividades humanas, a partir dos dados coletados por sensores de acelerômetro e giroscópio de smartphones. O dataset utilizado, *Human Activity Recognition Using Smartphones*, contém medições de 561 variáveis, coletadas de 30 voluntários enquanto realizavam seis atividades diárias distintas, como caminhar, subir escadas e ficar em pé.

Inicialmente, foi realizada uma **análise exploratória dos dados (EDA)**, na qual se investigou a distribuição das variáveis, as correlações entre elas e a necessidade de pré-processamento, como a **normalização** das variáveis, visto que os dados de acelerômetro e giroscópio apresentam escalas diferentes. Para reduzir a complexidade dos dados e facilitar a visualização dos padrões, foi aplicada a técnica de **Análise de Componentes Principais (PCA)**, que permitiu observar a distribuição das atividades em um espaço de duas dimensões.

O passo seguinte foi a implementação do algoritmo **K-means**, utilizando a biblioteca **Scikit-learn**, com o objetivo de agrupar as atividades de forma eficiente. A escolha do número ideal de clusters (K) foi realizada utilizando dois métodos: o **Método do Cotovelo** e o **Silhouette Score**. O primeiro método ajudou a identificar o ponto de inflexão na redução da inércia, enquanto o segundo forneceu uma medida da qualidade dos clusters gerados. Com base nessas análises, foi determinado que o número de clusters ideal seria **6**, correspondendo às 6 atividades no dataset.

O modelo K-means foi treinado com o número de clusters escolhido e os resultados avaliados com métricas como **inércia** e **Silhouette Score**. O modelo apresentou uma boa coesão entre os clusters, com um **Silhouette Score** de 0.61, indicando uma separação razoável entre as atividades. No entanto, algumas atividades, como "subir escadas" e "ficar em pé", apresentaram uma sobreposição maior, o que sugere que o agrupamento poderia ser aprimorado.

Este trabalho demonstrou a aplicabilidade do K-means para o agrupamento de atividades humanas em dados de sensores de smartphones, destacando a importância da escolha do número de clusters e da normalização dos dados. A análise dos resultados sugeriu que o modelo pode ser melhorado com a exploração de outros algoritmos de agrupamento e técnicas de redução de dimensionalidade.

Metodologia:

O desenvolvimento do projeto seguiu as seguintes etapas:

1. **Análise Exploratória dos Dados (EDA)**:
    - O dataset foi carregado e inspecionado para identificar características como distribuição das variáveis e possíveis correlações entre elas.

    - Foi realizada uma visualização inicial dos dados utilizando **PCA (Análise de Componentes Principais)** para reduzir a dimensionalidade e obter uma visualização clara das atividades.

2. **Pré-processamento e Normalização**:
    - As variáveis foram normalizadas utilizando o **StandardScaler**, para que todas tivessem a mesma escala, o que é essencial para o algoritmo K-means, pois ele depende de distâncias euclidianas.

3. **Implementação do K-means**:
    - O algoritmo K-means foi implementado utilizando a biblioteca **Scikit-learn**. Para garantir a convergência mais rápida e eficiente, utilizamos a inicialização **K-means++**.

    - O número ideal de clusters (K) foi escolhido com base em dois métodos: o **Método do Cotovelo** e o **Silhouette Score**. O primeiro analisa a inércia para diferentes valores de K, enquanto o segundo avalia a qualidade da separação entre os clusters.

4. **Avaliação dos Resultados**:
    - As métricas utilizadas para avaliar a qualidade dos clusters foram **inércia** (soma das distâncias quadradas dos pontos aos seus centroides) e **Silhouette Score** (uma medida da separação e coesão dos clusters).

Principais Resultados:

1. **Escolha do Número de Clusters**:
    - Através do **Método do Cotovelo**, foi possível observar que o ponto de inflexão ocorre para **K=6**, o que sugere que o número ideal de clusters seria 6, correspondente às 6 atividades presentes no dataset.
    - O **Silhouette Score** para K=6 foi próximo de 0.60, indicando uma separação razoável entre os clusters.

2. **Métricas de Avaliação**:
    - **Inércia**: A inércia do modelo para K=6 foi de 2.5e+06, o que indica uma boa coesão dos clusters.
    - **Silhouette Score**: O **Silhouette Score** para K=6 foi de 0.61, sugerindo que os clusters estão bem definidos, mas ainda pode haver algumas sobreposições entre as atividades.

3. **Visualização dos Clusters**: A redução de dimensionalidade com PCA permitiu a visualização dos clusters em 2D. A seguir, são apresentados os gráficos de dispersão que mostram os clusters identificados pelo K-means:

   (Inserir gráfico da dispersão dos clusters gerados pelo K-means)

4. **Análise dos Clusters**: A análise dos clusters revelou que as atividades mais dinâmicas, como caminhar e correr, foram agrupadas corretamente, enquanto atividades mais estáticas, como sentar e deitar, também formaram grupos distintos. No entanto, atividades como subir escadas e ficar em pé apresentaram um certo grau de sobreposição nos clusters, indicando uma possível confusão na interpretação dos dados.

Discussão:

A análise dos resultados revela que o modelo K-means conseguiu identificar corretamente a maioria das atividades, mas com algumas limitações. A separação das atividades "subir escadas" e "ficar em pé" não foi tão clara quanto as demais, o que pode indicar que os dados desses dois grupos possuem características semelhantes ou que a redução de dimensionalidade não foi suficientemente eficaz para capturar a diferença entre elas.

Limitações:

- O K-means, sendo um algoritmo baseado em distância, pode não ser o mais adequado para dados com formas complexas ou não lineares.

- A escolha do número de clusters foi baseada em métodos heurísticos, o que pode não refletir a realidade das atividades. Outras técnicas, como **DBSCAN** ou **Gaussian**

Mixture Models, poderiam ser exploradas para verificar se conseguem capturar melhor as relações entre as atividades.

Impacto das Escolhas:

- A normalização das variáveis foi crucial para o bom desempenho do K-means, já que as escalas das variáveis de acelerômetro e giroscópio eram muito diferentes.
- A escolha do número de clusters baseada nos gráficos do método do cotovelo e do silhouette score se mostrou eficaz, mas uma validação adicional com outras métricas de avaliação poderia aumentar a robustez do modelo.

Conclusão e Trabalhos Futuros:

Conclusões: O projeto demonstrou a eficácia do K-means para agrupar atividades humanas com base em dados de sensores. A escolha do número de clusters e a análise das métricas de avaliação indicam que o modelo conseguiu identificar as atividades com um bom nível de precisão, embora haja espaço para melhorias.

Trabalhos Futuros:

- Investigar a aplicação de outros algoritmos de agrupamento, como **DBSCAN** ou **Gaussian Mixture Models**, que podem ser mais adequados para dados não lineares.
- Explorar técnicas de **aumento de dados** para melhorar a robustez do modelo, especialmente para atividades com menos representatividade.
- Testar diferentes técnicas de redução de dimensionalidade (como t-SNE ou LLE) para melhorar a visualização e a separação dos clusters.

Referências:

1. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). *A Public Domain Dataset for Human Activity Recognition Using Smartphones*. ESANN 2013.

2. UCI Machine Learning Repository - Human Activity Recognition Using Smartphones. Disponível em: https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones.

3. Scikit-learn Documentation. Disponível em: https://scikit-learn.org/.