Clusterização de Atividades Humanas com K-means

Nome dos Alunos:
Onécio Araujo Ribeiro
Janailton Ferreira da Costa

23 de novembro de 2024

Resumo:

Este projeto emprega o algoritmo K-means para agrupar atividades humanas, utilizando dados sensoriais do conjunto de dados "Human Activity Recognition Using Smartphones". O processo envolveu a análise exploratória dos dados, seguido pela normalização, definição do número ideal de clusters empregando os métodos do cotovelo e Silhouette Score, além da visualização dos resultados em um espaço de duas dimensões por meio da Análise de Componentes Principais (PCA). Os achados demostraram uma clara separação entre os clusters, sugerindo que os sensores capturam padrões distintos nas atividades humanas.

Introdução

O reconhecimento de atividades humanas é fundamental em áreas como saúde, esportes e tecnologias para dispositivos inteligentes. Dados sensoriais obtidos de acelerômetros e giroscópios em smartphones possibilitam a identificação de padrões em diversas atividades, por exemplo, caminhar, subir escadas e ficar em pé. A utilização de métodos de clusterização não supervisionada, como o K-means, oferece uma maneira eficaz de explorar padrões em dados complexos. Este projeto é motivado pela demanda por soluções que permitam identificar atividades humanas sem supervisão direta, aplicando técnicas que sejam ao mesmo tempo escaláveis e fáceis de interpretar.

Metodologia

1. Análise Exploratória de Dados (EDA):

- Carregamento do dataset test.csv.
- Inspeção de valores ausentes, estatísticas descritivas e análise de correlações entre variáveis.
- Visualizações, incluindo histogramas e mapa de calor, para entender a distribuição e a relação entre os dados.

2. Pré-processamento:

- **Normalização:** Os dados foram normalizados com StandardScaler para equilibrar a influência das variáveis sensoriais.
- **Redução de dimensionalidade (PCA):** Foi utilizada para facilitar a interpretação e visualização dos clusters.

3. Algoritmo de Clusterização:

- **Implementação do K-means:**
    - Inicialização com K-means++ para melhor convergência.
    - Avaliação do número ideal de clusters utilizando:
        - Método do cotovelo: Identificação visual do ponto de inflexão na inércia.
        - Silhouette Score: Métrica para avaliar a coesão dos clusters.

4. Visualização e Avaliação:

- Os clusters foram visualizados em 2D usando PCA, e métricas como inércia e Silhouette Score foram registradas para avaliar a qualidade dos clusters.

Resultados

1. **Número Ótimo de Clusters:**

O método do cotovelo indicou o ponto de inflexão em k=3k=3k=3, corroborado pelo Silhouette Score, que alcançou um valor satisfatório para essa configuração.

2. Avaliação dos Clusters:

- **Inércia Final:**
- **Silhouette Score:**

3. Visualizações:

- Mapa bidimensional dos clusters usando PCA revelou uma separação clara entre os grupos, indicando que o algoritmo conseguiu identificar padrões distintos.

Discussão

Os resultados mostram que o K-means é uma ferramenta eficaz para identificar padrões de atividades humanas com base em dados sensoriais. No entanto, a abordagem apresenta algumas limitações:

- **Natureza Não Supervisionada:** Sem rótulos, a interpretação dos clusters depende de correlações inferidas.
- **Dimensionalidade Reduzida:** A redução via PCA pode causar perda de informações em componentes descartados.
- **Balanceamento de Dados:** Variáveis com maior variância tendem a dominar a clusterização, justificando a normalização.

Apesar disso, a separação clara dos clusters sugere que o K-means capturou informações relevantes sobre os padrões de atividade.

Conclusão e Trabalhos Futuros

Conclusão:

O projeto demonstrou o uso bem-sucedido do K-means para clusterizar atividades humanas com base em dados sensoriais. A análise visual e as métricas quantitativas corroboram a qualidade dos clusters.

Trabalhos Futuros:

- Explorar métodos supervisionados para comparar os clusters com rótulos de atividade reais.
- Testar outros algoritmos de clusterização, como DBSCAN e Mean Shift, para avaliar ganhos de qualidade.
- Integrar variáveis temporais para capturar a sequência de atividades.

Referências

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

Python Software Foundation. (2024). Python Language Reference, version 3.10.