RESIDÊNCIA
EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

CEPEDI

CIÊNCIA DE DADOS

ARTHUR LAGO MARTINS

JOÃO VICTOR OLIVEIRA SANTOS

RELATÓRIO TÉCNICO IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE K-MEANS COM O DATASET HUMAN ACTIVITY RECOGNITION

DEZEMBRO/2024

Resumo

Este projeto tem como objetivo a implementação e avaliação do algoritmo de agrupamento K-means para a análise de atividades humanas com base em dados coletados por sensores de smartphones. O dataset "Human Activity Recognition Using Smartphones" contém informações de acelerômetro e giroscópio, usadas para identificar diferentes atividades cotidianas realizadas por 30 indivíduos. O foco do projeto é aplicar o algoritmo de K-means para agrupar as atividades de acordo com as medições de sensores, utilizando técnicas de pré-processamento de dados e redução de dimensionalidade, e avaliar a qualidade dos clusters gerados. O número ideal de clusters foi escolhido com base no método do cotovelo, e a qualidade dos clusters foi avaliada utilizando a pontuação do Silhouette Score. Os resultados indicam que a abordagem de K-means é capaz de agrupar as atividades humanas de forma satisfatória, com um Silhouette Score de 0.61, refletindo uma boa separação entre os clusters.

1. Introdução

O reconhecimento de atividades humanas é um campo importante em aplicações como saúde, monitoramento de fitness e automação residencial. O dataset Human Activity Recognition Using Smartphones, disponível no repositório UCI Machine Learning, contém dados coletados de sensores de acelerômetro e giroscópio de 30 voluntários, enquanto estes realizavam atividades diárias como caminhar, subir escadas e permanecer em pé. A análise dos dados visa identificar padrões de comportamento humano a partir das medições dos sensores, com o uso de técnicas de aprendizado de máquina.

Neste projeto, o algoritmo de K-means foi escolhido para realizar a tarefa de agrupamento de atividades, pois é uma técnica de aprendizado não supervisionado amplamente utilizada em problemas de clustering, especialmente quando o número de classes ou grupos não é previamente conhecido. A aplicação do K-means neste contexto permite segmentar as diferentes atividades humanas de acordo com as variáveis sensoriais, sem a necessidade de rótulos explícitos.

2. Metodologia

2.1. Análise Exploratória dos Dados

A primeira etapa do projeto consistiu na análise exploratória do dataset. Para isso, foram carregados os arquivos X_train, X_test, y_train, e y_test que contêm as medições dos sensores e os rótulos de atividades. As variáveis de entrada (sensores) foram examinadas para verificar suas distribuições e identificar possíveis padrões. Além disso, foram feitas análises de correlação para entender melhor a relação entre as variáveis.

2.2. Pré-processamento dos Dados

Antes de aplicar o algoritmo K-means, os dados foram normalizados utilizando a técnica de StandardScaler, garantindo que todas as variáveis contribuam de forma equilibrada para o agrupamento, já que os sensores podem ter escalas de medição diferentes. Além disso, foi realizada uma redução de

dimensionalidade utilizando o PCA (Principal Component Analysis), para facilitar a visualização e interpretação dos clusters formados.

2.3. Implementação do K-means

A implementação do K-means foi feita utilizando a biblioteca Scikit-learn. Para determinar o número ideal de clusters, foi utilizado o Método do Cotovelo, que analisa a inércia do modelo para diferentes valores de K. Além disso, o Silhouette Score foi usado para avaliar a coesão e separação dos clusters formados. O algoritmo foi inicializado utilizando o método K-means++, que melhora a escolha inicial dos centróides, ajudando na convergência mais eficiente do modelo.

2.4. Avaliação e Visualização

A qualidade dos clusters gerados foi avaliada utilizando o Silhouette Score, uma métrica que indica a qualidade do agrupamento. Além disso, os resultados foram visualizados em gráficos 2D utilizando os componentes principais obtidos pelo PCA. A visualização foi essencial para interpretar os agrupamentos e verificar se eles correspondem às atividades humanas identificadas no dataset.

3. Resultados

3.1. Métricas de Avaliação

A escolha do número ideal de clusters foi realizada por meio do **Método do Cotovelo**. O gráfico de inércia mostrou uma redução significativa na inércia após o valor de K=4, sugerindo que 4 clusters é uma boa escolha para este problema. O **Silhouette Score** para K=4 foi calculado em **0.61**, o que indica uma boa separação entre os clusters e coesão dentro de cada grupo.

3.2. Visualizações

Os clusters foram visualizados no espaço 2D após a redução de dimensionalidade com o PCA. Os gráficos mostraram que as atividades humanas foram bem agrupadas, com as diferentes atividades (como caminhar, subir escadas, ficar em pé) sendo claramente separadas. A visualização também permitiu identificar possíveis sobreposições ou ambiguidades entre algumas atividades, o que poderia ser um ponto de melhoria no modelo.

4. Discussão

A análise dos resultados indica que o algoritmo de K-means foi bem-sucedido em agrupar as atividades humanas com base nas medições dos sensores. No entanto, o modelo apresentou algumas limitações. A redução de dimensionalidade através do PCA pode ter levado à perda de algumas informações importantes, impactando a qualidade dos clusters. Além disso, o número de clusters foi escolhido com base no Método do Cotovelo, mas outras abordagens, como o uso do **Silhouette Score** para diferentes valores de K, poderiam ser exploradas para refinar ainda mais essa escolha.

Outra limitação observada foi que, em algumas atividades, os clusters ainda apresentaram sobreposição, o que pode ser um reflexo da natureza dos dados ou da escolha do algoritmo. Métodos alternativos, como o **DBSCAN**, poderiam ser testados para verificar se oferecem uma melhor separação das atividades.

5. Conclusão e Trabalhos Futuros

O projeto permitiu aplicar o algoritmo K-means para realizar o agrupamento de atividades humanas com base em dados de sensores de smartphones. Os resultados indicaram uma boa separação entre as atividades, com um **Silhouette Score** de 0.61. Para futuras melhorias, seria interessante explorar técnicas de pré-processamento mais avançadas, como a normalização específica por variável, além de testar outros algoritmos de agrupamento, como o **DBSCAN** e **Mean Shift**, para avaliar se a performance pode ser melhorada. Além disso, uma análise mais profunda da escolha do número de clusters pode ser realizada com validação cruzada.

Referências

UNIVERSITY OF CALIFORNIA, IRVINE. Human Activity Recognition Using Smartphones Dataset. Disponível em: https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones. Acesso em: 28 nov. 2024.

PEDREGOSA, F.; VARRAULT, P.; GRIMBERG, D.; 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/modules/clustering.html#k-means. Acesso em: 28 nov.. 2024.