Relatório Técnico: Implementação e Análise do Algoritmo de K-Means

Álisson Natan dos Anjos Pinheiro
Larissa Ribeiro Firminio

29/11/2024

Resumo

O objetivo deste projeto foi aplicar o algoritmo de K-means para realizar a análise de agrupamento de dados do dataset Human Activity Recognition using Smartphones. O projeto incluiu etapas de análise exploratória, implementação do modelo de K-means, e avaliação da qualidade dos clusters obtidos.

Durante o processo, foi realizada a escolha do número de clusters usando métodos como o método do cotovelo e o silhouette score. O modelo foi otimizado por meio de técnicas como normalização dos dados e inicialização K-means++. Como resultado, observou-se uma boa separação dos clusters com uma inércia reduzida e um silhouette score satisfatório.

Introdução

O reconhecimento de atividades humanas a partir de dados coletados por smartphones tem se tornado uma área de interesse crescente em diversas áreas, como a saúde e a automação.

O dataset Human Activity Recognition using Smartphones contém dados de sensores (acelerômetro e giroscópio) de 30 voluntários realizando atividades diárias como caminhar, subir escadas e ficar em pé. O objetivo deste projeto foi usar o algoritmo de K-means para agrupar esses dados e identificar padrões que representem diferentes atividades humanas.

A escolha do K-means justifica-se pelo fato de ser um algoritmo simples e eficiente para tarefas de agrupamento sem supervisão, permitindo identificar automaticamente os padrões nos dados.

Metodologia

A metodologia do projeto foi dividida em algumas etapas, sendo elas as etapas: Análise Exploratória dos Dados, Implementação do Algoritmo de K-means, Normalização e Repetição do Modelo, e Avaliação e Métricas.

1. Análise Exploratória dos Dados

A análise exploratória foi realizada para entender melhor o dataset e suas características. As principais etapas foram:

- Distribuição das variáveis: Análise de variáveis usando gráficos de boxplot e histogramas;
- Correlação entre variáveis: Construção de uma matriz de correlação para identificar variáveis altamente correlacionadas e selecionar aquelas mais relevantes para o agrupamento;

- Redução de dimensionalidade: Utilização do Principal Component Analysis para reduzir a dimensionalidade dos dados e facilitar a visualização dos clusters.

2. Implementação do Algoritmo de K-means

Foi implementado o algoritmo de K-means com o K-means++ para inicialização eficiente dos centróides e otimização do processo de convergência. O número de clusters foi inicialmente estimado utilizando o Elbow Method e o Silhouette Score, para determinar o valor do número de clusters que melhor representava os agrupamentos das atividades.

3. Normalização e Repetição do Modelo

Devido à alta variabilidade nas escalas dos sensores, foi realizada a normalização dos dados com o StandardScaler.

Ademais, o modelo de K-means foi executado múltiplas vezes para verificar a consistência dos clusters formados, analisando a estabilidade e a confiabilidade dos resultados.

4. Avaliação e Métricas

Já quando referente as métricas de avaliação utilizadas, foram usadas as seguintes:

- Inércia: A qual mediu a soma das distâncias quadradas entre os pontos e seus centróides;
- Silhouette Score: A qual indicou a qualidade dos clusters, avaliando a coesão e a separação entre eles.

Ao fim dessa etapa, o resultado obtido após a execução do algoritmo foi um valor de inércia de 881007.9475 e um valor de silhouette de 0.4842.

Essas métricas indicam que os clusters obtidos têm boa coesão interna, ou seja, eles possuem alta similaridade entre os pontos dentro de um cluster. Além disso, também há uma boa separação entre os clusters.

Ademais, quando referente aos gráficos de visualização, eles mostraram que a redução de dimensionalidade usando o Principal Component Analysis foi eficaz para representar visualmente a separação dos clusters. O gráfico de 2D e o gráfico de 3D mostraram claramente a distribuição dos dados nos clusters formados pelo K-means.

Discussão

Embora os resultados tenham sido satisfatórios, algumas limitações foram observadas. Entre essas limitações, está o fato de que a escolha do número de clusters pode não ser perfeita para todos os contextos, já que a definição desse valor é um desafio, especialmente quando se trata de problemas não supervisionados. Além disso, também há o fato de que a qualidade do modelo pode ser impactada pela qualidade dos dados, como a possível presença de outliers ou valores faltantes.

A escolha de K-means foi apropriada devido à sua simplicidade e boa performance em problemas de clustering, mas, em casos mais complexos, seria interessante explorar outros algoritmos como DBSCAN ou Hierarchical Clustering para comparar os resultados.

Conclusão e Trabalhos Futuros

Este projeto demonstrou a eficácia do algoritmo de K-means no agrupamento de dados de reconhecimento de atividades humanas. As etapas de análise exploratória, escolha do número de clusters, e avaliação de desempenho garantiram a qualidade dos resultados obtidos. A normalização e a inicialização com K-means++ foram essenciais para otimizar o processo de agrupamento.

Para melhorar os resultados, é possível explorar outros métodos de clustering como o Hierarchical Clustering ou o DBSCAN para verificar se eles conseguem resultados melhores que os obtidos a partir do método utilizado neste trabalho. Além disso, também é possível incluir mais variáveis no modelo, como informações temporais ou contextuais, as quais podem ser úteis para melhorar a precisão da classificação das atividades.

Referências

UCI Machine Learning Repository - Human Activity Recognition Using Smartphones Dataset. Disponível em: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Scikit-learn Documentation. Disponível em: https://scikit-learn.org/stable/documentation.html