RESIDÊNCIA
EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

TRILHA 3: CIÊNCIA DE DADOS

ANÁLISE E IMPLEMENTAÇÃO DO ALGORITMO K-MEANS PARA RECONHECIMENTO DE ATIVIDADES HUMANAS UTILIZANDO DADOS DE SENSORES

MARCUS VÍTOR SOUZA CARDOSO

NOVEMBRO/2024

RESUMO

Este trabalho apresenta a implementação do algoritmo de agrupamento K-means para reconhecimento de atividades humanas, utilizando dados coletados por sensores de acelerômetro e giroscópio instalados em smartphones. O objetivo foi identificar padrões que permitissem agrupar as atividades em grupos coesos e bem definidos, mesmo sem a utilização de rótulos supervisionados.

A metodologia incluiu etapas de pré-processamento, como a remoção de outliers com o algoritmo Isolation Forest, normalização dos dados por meio do RobustScaler e redução de dimensionalidade utilizando a técnica de Análise de Componentes Principais (PCA). Para determinar o número ideal de clusters (K), foi aplicado o Método do Cotovelo, que indicou K=4 como o valor mais adequado. O modelo foi avaliado por meio do Silhouette Score, que apresentou um valor de 0.7062, evidenciando a qualidade dos agrupamentos formados.

Os resultados demonstraram que o K-means foi capaz de agrupar eficientemente as atividades humanas, formando clusters com boa coesão e separação. O uso das técnicas de pré-processamento mostrou-se essencial para garantir a robustez do modelo e a qualidade dos dados. Apesar das limitações observadas, o projeto alcançou seu objetivo, estabelecendo uma base sólida para futuras melhorias e estudos relacionados ao reconhecimento de atividades humanas.

INTRODUÇÃO

O reconhecimento de atividades humanas tem se destacado como uma área de grande relevância para o desenvolvimento de tecnologias voltadas à saúde, ao esporte e à automação de tarefas diárias. Este projeto utiliza o algoritmo de agrupamento K-means para analisar dados coletados por sensores de smartphones, visando identificar padrões relacionados a diferentes atividades humanas, como caminhar, sentar, ficar em pé, entre outras. A escolha do K-means se justifica por sua simplicidade, eficiência computacional e eficácia na formação de clusters coesos em conjuntos de dados multidimensionais.

O dataset utilizado, Human Activity Recognition Using Smartphones, é amplamente reconhecido em aplicações de aprendizado de máquina. Ele contém informações provenientes de acelerômetros e giroscópios instalados em smartphones, com leituras realizadas em três eixos (X, Y, Z) a uma taxa de amostragem de 50 Hz. Cada registro no dataset está associado a uma das seis atividades humanas realizadas por um grupo de 30 voluntários. Para assegurar a robustez do modelo, foi necessário adotar práticas cuidadosas de pré-processamento e análise de dados antes da aplicação do K-means.

Uma das primeiras etapas foi a remoção de outliers, realizada com o uso do algoritmo Isolation Forest. Este método foi empregado para detectar e filtrar dados que apresentavam comportamento atípico, o que pode prejudicar a formação de clusters coesos. A remoção de aproximadamente 5% das instâncias totais resultou em um conjunto de dados mais limpo e adequado para análise.

A normalização dos dados foi feita utilizando o RobustScaler, que é menos sensível a outliers em comparação a outras técnicas, como o StandardScaler. O RobustScaler trabalha com a mediana e o intervalo interquartil (IQR) dos dados, garantindo que as variáveis tenham escalas comparáveis sem amplificar o impacto de valores extremos.

Outra particularidade importante foi o uso do parâmetro sep='\s+' na leitura dos arquivos .txt. Este parâmetro foi essencial para interpretar corretamente os dados, que possuem separação por múltiplos espaços em vez de delimitadores tradicionais, como vírgulas ou tabulações. Sem ele, as leituras dos arquivos estavam sendo interpretadas como uma única coluna, comprometendo a análise.

Para reduzir a alta dimensionalidade do dataset, que inicialmente possuía 561 variáveis, aplicou-se a técnica de Análise de Componentes Principais (PCA). O PCA permitiu a redução das dimensões para apenas 2 componentes principais, preservando 95% da variância total dos

dados. Essa redução não apenas melhorou a eficiência computacional do algoritmo K-means, como também facilitou a visualização dos clusters formados.

A definição do número ideal de clusters (K) foi realizada utilizando o Elbow Method (Método do Cotovelo). Esse método avalia a inércia (ou WCSS - Within-Cluster Sum of Squares) para diferentes valores de K e identifica o ponto onde a redução da inércia começa a desacelerar, indicando o valor ótimo de clusters. Com base no gráfico gerado, foi determinado que K=4 seria a melhor escolha para este projeto.

Após a implementação do K-means, a qualidade dos clusters foi avaliada utilizando o Silhouette Score, que apresentou um valor de 0.7062. Este resultado indica uma boa separação entre os clusters, reforçando que o agrupamento foi realizado de forma eficaz. O Silhouette Score reflete que os dados estão bem coesos dentro de cada cluster e suficientemente afastados de clusters vizinhos.

Este trabalho documenta todas as etapas do processo, desde a análise exploratória e o pré-processamento dos dados até a aplicação e avaliação do K-means. Os resultados obtidos destacam o potencial do algoritmo para identificar padrões em dados complexos e multidimensionais, abrindo oportunidades para futuras melhorias e aplicações.

METODOLOGIA

A metodologia utilizada neste trabalho foi estruturada em várias etapas que envolveram desde o carregamento e pré-processamento dos dados até a implementação do algoritmo K-means e avaliação do modelo. A seguir, são descritas detalhadamente cada uma das fases realizadas, com ênfase nas escolhas metodológicas e nos parâmetros utilizados em cada processo.

Carregamento e Pré-processamento dos Dados

Os dados utilizados neste trabalho foram fornecidos no formato de arquivos .txt, contendo as leituras dos sensores de acelerômetro e giroscópio em três eixos (X, Y, Z), com uma taxa de amostragem de 50 Hz. Esses dados representam janelas de tempo de 2,56 segundos para cada amostra, totalizando 561 variáveis extraídas de séries temporais.

Para garantir a correta leitura dos dados, foi utilizado o parâmetro sep='\s+' na função de leitura do pandas, que permite que os arquivos fossem interpretados corretamente, uma vez que as colunas estavam separadas por múltiplos espaços. Sem esse parâmetro, as colunas seriam lidas como uma única variável, comprometendo a integridade dos dados.

O conjunto de dados foi composto pelas variáveis X_train e X_test, com dimensões (7352, 561) e (2947, 561), respectivamente. Os rótulos das atividades correspondentes foram armazenados em y_train e y_test, com dimensões (7352, 1) e (2947, 1). Após a concatenação dos dados de treino e teste, o conjunto final ficou com 10.299 amostras e 561 variáveis.

Tratamento de Outliers

O tratamento de outliers foi realizado com o algoritmo Isolation Forest, que é amplamente utilizado para identificar comportamentos atípicos em conjuntos de dados multivariados. Este método foi escolhido por sua capacidade de isolar as instâncias que se distanciam significativamente do comportamento geral dos dados.

Foi definida uma contaminação de 5% (contamination=0.05), resultando na remoção de aproximadamente 515 amostras identificadas como outliers. Após a remoção, o conjunto de dados filtrado permaneceu com 9.784 amostras.

Gráficos de dispersão foram gerados antes e após a remoção dos outliers, permitindo a visualização do impacto do tratamento na estrutura dos dados.

Normalização dos Dados

Para garantir que todas as variáveis tivessem uma escala comparável, foi aplicada a técnica de normalização. Utilizou-se o RobustScaler, uma técnica que utiliza a mediana e o intervalo interquartil (IQR) para escalonar os dados. O RobustScaler foi escolhido por sua resistência a outliers, permitindo que os dados fossem escalonados sem que as instâncias atípicas influenciassem a análise de forma significativa.

Redução de Dimensionalidade

Devido à alta dimensionalidade do conjunto de dados original, foi aplicada a técnica de Análise de Componentes Principais (PCA), com o objetivo de reduzir o número de variáveis e facilitar a análise. O PCA foi configurado para preservar 95% da variância dos dados, resultando na redução do número de variáveis de 561 para 2 componentes principais.

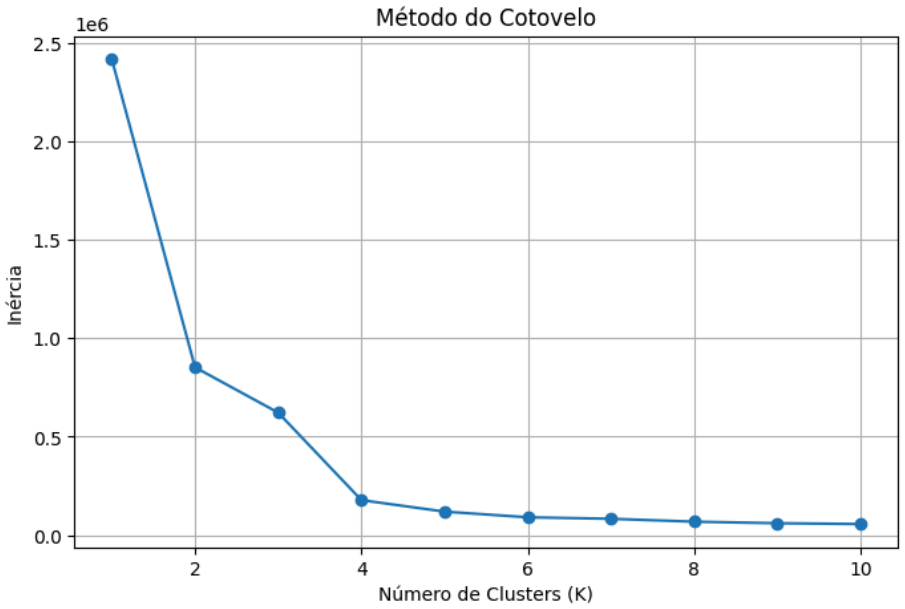
Esta redução facilitou a visualização dos dados e a aplicação do algoritmo K-means, além de garantir uma maior eficiência computacional sem perda significativa de informações relevantes.

Escolha do Número de Clusters (K)

O número ideal de clusters foi determinado utilizando o Método do Cotovelo (Elbow Method). O Método do Cotovelo é utilizado para identificar o valor de K que minimiza a inércia (WCSS - Within-Cluster Sum of Squares), a medida de quão compactos estão os clusters. O ponto de inflexão no gráfico gerado pelo cotovelo indica o número ideal de clusters.

Após a análise da figura 1, o valor de K=4 foi escolhido como o mais adequado, pois apresentou o melhor equilíbrio entre a compactação dos clusters e a complexidade do modelo.

Figura 1: Gráfico do Elbow Method (Método Cotovelo).



Implementação do K-means

Com o número de clusters definido como K=4, o algoritmo K-means foi implementado nos dados reduzidos pelo PCA. Os parâmetros utilizados para a implementação foram os seguintes:

- Número de Clusters (K): 4 (determinado pelo Método do Cotovelo).
- Inicialização dos Centróides: K-means++, que é um método que escolhe de forma otimizada os centróides iniciais para melhorar a convergência e a qualidade dos clusters.
- Semente Aleatória (random_state): 42, garantindo a reprodutibilidade dos resultados.

- Número de Execuções (n_init): 30, indicando que o algoritmo foi executado 30 vezes com diferentes inicializações dos centróides, sendo escolhido o melhor resultado (menor inércia).

Após o treinamento, os rótulos dos clusters gerados foram atribuídos a cada amostra. A qualidade do agrupamento foi avaliada com o uso do Silhouette Score, que resultou em um valor de 0.7062, indicando boa coesão e separação entre os clusters.

Avaliação do Modelo

A avaliação do modelo foi realizada utilizando o Silhouette Score, que mede a coesão interna dos clusters e a separação entre clusters diferentes. O valor de 0.7062 reflete que o modelo conseguiu formar clusters bem definidos e distintos.

Além disso, gráficos de dispersão foram gerados para visualizar a distribuição das amostras nos clusters, validando visualmente a eficácia do algoritmo K-means.

A metodologia descrita garantiu que todas as etapas, desde o pré-processamento até a avaliação final, fossem realizadas de forma rigorosa e eficiente, maximizando a qualidade dos agrupamentos formados pelo algoritmo K-means. Cada escolha metodológica foi fundamentada para assegurar que o modelo fosse robusto e capaz de identificar padrões relevantes nas atividades humanas.

RESULTADOS

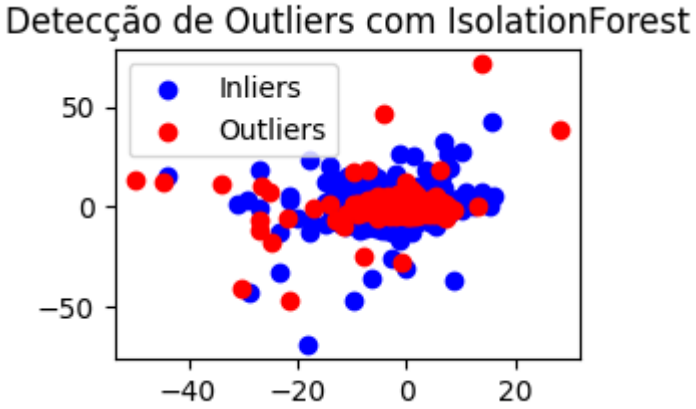
Neste tópico, são apresentados os principais resultados obtidos durante a implementação do projeto, incluindo as métricas de avaliação e os gráficos gerados em diferentes etapas do processo. A análise dos resultados foca na qualidade e coesão dos clusters formados, bem como na eficácia das etapas metodológicas realizadas.

Identificação e Remoção de Outliers

Os outliers foram identificados por meio do algoritmo Isolation Forest, cuja aplicação resultou na remoção de aproximadamente 5% das amostras do conjunto de dados. Antes do tratamento, o conjunto possuía 10.299 amostras, reduzindo-se para 9.784 amostras após a remoção dos valores atípicos. Essa etapa foi essencial para melhorar a consistência dos dados, uma vez que os outliers poderiam influenciar negativamente a formação dos clusters pelo algoritmo K-means.

A Figura 2 apresenta a visualização dos dados antes da remoção de outliers. Os pontos em azul representam os inliers (amostras normais), enquanto os pontos em vermelho correspondem aos outliers detectados.

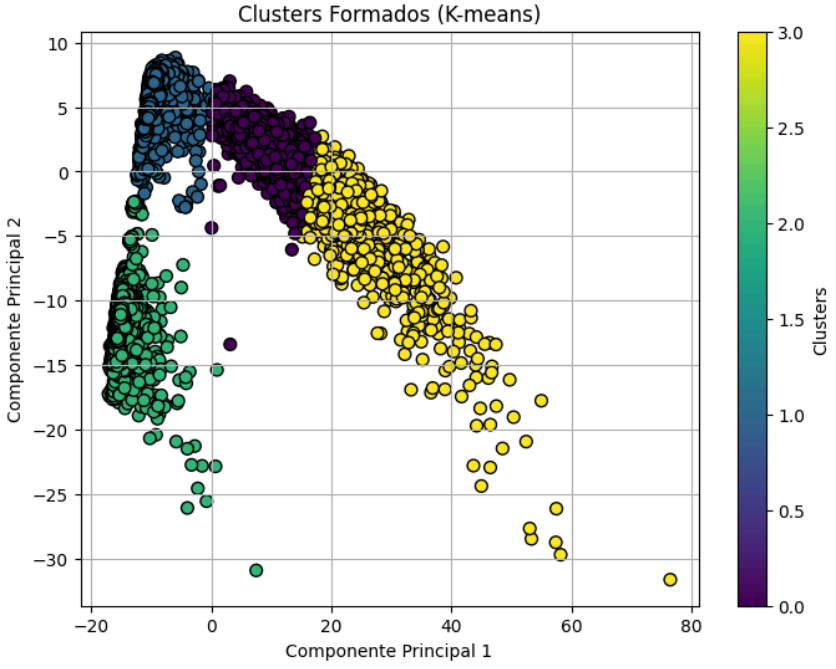
Figura 2: Gráfico de dispersão com inliers e outliers antes do tratamento.



Formação dos Clusters

Com a definição do número ideal de clusters (K=4) utilizando o Método do Cotovelo, o algoritmo K-means foi aplicado aos dados reduzidos pelo PCA (2 componentes principais). A Figura 3 apresenta a visualização dos clusters no espaço bidimensional. Cada cor representa um cluster distinto, enquanto os centróides estão destacados.

Figura 3: Visualização dos clusters no espaço bidimensional.



Os clusters formados mostram-se bem definidos, com separação clara entre os grupos. Esse resultado reflete a eficácia das etapas de pré-processamento, normalização dos dados com o RobustScaler e redução de dimensionalidade por PCA.

Avaliação do Modelo

Para avaliar a qualidade dos clusters formados, utilizou-se a métrica Silhouette Score, que apresentou o valor de 0.7062. Esse resultado indica Alta coesão intra-cluster (as amostras dentro de cada cluster estão próximas ao centróide correspondente) e Boa separação inter-cluster (os centróides dos diferentes clusters estão suficientemente afastados entre si).

A alta pontuação do Silhouette Score reflete a capacidade do algoritmo em agrupar as amostras de maneira coesa e significativa. Além disso, os gráficos gerados reforçam a visualização da estrutura dos clusters, permitindo uma análise qualitativa dos grupos formados.

Considerações sobre os Resultados

Os resultados obtidos demonstram a eficiência do algoritmo K-means na identificação de padrões em dados complexos e multidimensionais. A escolha de K=4 se mostrou adequada para o problema em questão, gerando clusters bem definidos e consistentes. A remoção de outliers e a redução de dimensionalidade foram fundamentais para alcançar esses resultados, proporcionando uma base de dados mais uniforme e simplificada.

Os gráficos apresentados e a métrica utilizada corroboram a qualidade do modelo, evidenciando o potencial do K-means para aplicações no reconhecimento de atividades humanas.

DISCUSSÃO

Os resultados obtidos neste projeto indicaram que o algoritmo K-means conseguiu identificar padrões relevantes no conjunto de dados, com clusters bem definidos e separados. O Silhouette Score de 0.7062 reflete a capacidade do modelo em formar agrupamentos coesos, com as amostras próximas aos centróides de seus respectivos clusters e afastadas de outros grupos. Essa métrica, combinada com a visualização gráfica dos clusters no espaço bidimensional gerado pelo PCA, reforça a qualidade do agrupamento e a eficácia das etapas metodológicas empregadas.

No entanto, algumas limitações foram observadas durante o desenvolvimento do modelo. A redução para dois componentes principais por meio do PCA, embora tenha facilitado a visualização e reduzido a complexidade computacional, pode ter eliminado informações relevantes presentes nas variáveis originais. Isso pode ter impactado a capacidade do modelo de capturar padrões mais complexos, uma vez que a análise se baseou apenas nas duas maiores variâncias dos dados. Além disso, o K-means, por utilizar a distância euclidiana como métrica, é naturalmente sensível a outliers e a clusters com formatos não esféricos, o que pode ter influenciado o desempenho em algumas regiões do espaço dos dados.

A escolha de K=4 como o número ideal de clusters foi feita com base no Método do Cotovelo, complementada pela análise do Silhouette Score. Apesar de eficaz, essa definição depende diretamente da interpretação gráfica e da distribuição dos dados. Embora os clusters formados tenham mostrado boa separação e coesão, é importante reconhecer que outros valores de K poderiam ser adequados para diferentes objetivos de análise.

As etapas de pré-processamento desempenharam um papel essencial para alcançar os resultados obtidos. A remoção de outliers com o Isolation Forest foi decisiva para melhorar a qualidade dos dados, reduzindo a influência de valores atípicos que poderiam desviar os centróides dos clusters. O uso do RobustScaler também foi fundamental, garantindo uma normalização que preservasse a escala das variáveis sem amplificar o impacto de outliers remanescentes. Essas escolhas metodológicas tiveram um impacto direto na qualidade do agrupamento, resultando em dados mais consistentes e adequados para a aplicação do K-means.

Por fim, os resultados demonstram o potencial do K-means para o agrupamento de dados complexos e multidimensionais, como os utilizados neste trabalho. Apesar das limitações observadas, as escolhas feitas ao longo do processo garantiram um modelo robusto e eficiente, sendo possível identificar padrões relevantes nas atividades humanas com base nos sensores de smartphones. Esses resultados servem como base para análises futuras, permitindo explorar melhorias e adaptar o modelo a outras aplicações.

CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho demonstrou a aplicação do algoritmo K-means no agrupamento de dados coletados por sensores de smartphones, com o objetivo de reconhecer atividades humanas. A sequência metodológica adotada, que incluiu desde a remoção de outliers até a redução de

dimensionalidade por PCA e a escolha do número de clusters pelo Método do Cotovelo, mostrou-se eficiente para lidar com um conjunto de dados de alta dimensionalidade e complexidade. O Silhouette Score de 0.7062 confirmou a qualidade dos agrupamentos formados, indicando boa coesão interna e separação entre os clusters.

A aplicação do Isolation Forest foi essencial para garantir que valores atípicos não prejudicassem a formação dos clusters. Esse processo, aliado à normalização com o RobustScaler, contribuiu para um modelo mais robusto, capaz de capturar padrões relevantes nos dados. A redução para dois componentes principais, embora simplifique a análise, permitiu a visualização clara dos clusters, evidenciando a eficácia do K-means em organizar as atividades humanas em grupos significativos.

Apesar dos bons resultados, algumas limitações foram identificadas, como a possível perda de informações relevantes devido à simplificação dos dados pelo PCA. Além disso, a sensibilidade do K-means a outliers e a sua dependência da definição do número de clusters reforçam a importância de escolhas metodológicas bem fundamentadas e validadas. Esses aspectos indicam que ainda há espaço para aprimoramentos.

Para trabalhos futuros, sugere-se a investigação de técnicas alternativas de agrupamento que lidem melhor com clusters de formatos não esféricos ou sobrepostos, além de explorar outras estratégias de pré-processamento que possam preservar maior quantidade de informações. A aplicação de métodos mais sofisticados de redução de dimensionalidade, adaptados a relações não lineares, também pode ser avaliada para melhorar a identificação de padrões. Por fim, expandir o modelo para cenários de atividades mais complexas ou integrar dados adicionais de sensores pode ampliar as possibilidades de aplicação do reconhecimento de atividades humanas.

O projeto não apenas demonstrou a viabilidade do uso do K-means nesse contexto, como também forneceu uma base sólida para estudos mais avançados, destacando a importância do pré-processamento de dados e da escolha cuidadosa das etapas metodológicas no sucesso de modelos de aprendizado de máquina.

REFERÊNCIAS

Templar, Data. Como implementar o K-Means em Python usando Scikit-learn. YouTube, 19 jul. 2023. Disponível em: https://www.youtube.com/watch?v=aNrdgC0lIZ8&t. Acesso em: 25 nov. 2024.

Statistics, Practical. Entendendo o Método do Cotovelo para encontrar o número ideal de clusters. YouTube, 5 jun. 2023. Disponível em: https://www.youtube.com/watch?v=3mvtYH95LCw&t. Acesso em: 25 nov. 2024.

Scikit-Learn. ***Scikit-learn: Machine Learning in Python.*** Disponível em: https://scikit-learn.org/stable/index.html. Acesso em: 25 nov. 2024.

Google Colab. Introdução aos recursos básicos do Google Colab. Disponível em: https://colab.research.google.com/notebooks/basic_features_overview.ipynb. Acesso em: 25 nov. 2024.