RESIDÊNCIA EM TIC 36

Guilherme Melo dos Santos, Gleidson de Meireles Costa

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

Feira de Santana

01/12/2024

Resumo: Este trabalho investiga o uso do algoritmo **K-Means** para análise do **UCI HAR Dataset**, composto por 561 **features** de sensores de smartphones coletadas durante seis atividades realizadas por 30 voluntários. O objetivo é avaliar a capacidade do **K-Means** de identificar padrões nos dados sem supervisão, explorando seu potencial em cenários com rótulos limitados. Os dados foram normalizados e reduzidos em dimensionalidade com PCA, e o número de clusters foi estimado pelo método do cotovelo. A qualidade dos agrupamentos foi avaliada com métricas como **Silhouette Score** e índice Rand ajustado. Os resultados mostram que os agrupamentos coincidem significativamente com as atividades reais em cenários de movimentos distintos, demonstrando a viabilidade do **K-Means** como ferramenta preliminar para reconhecimento de atividades humanas.

1. Introdução

O reconhecimento de atividades humanas (HAR - Human Activity Recognition) utilizando dados de sensores inerciais tem se tornado um tema central em aplicações de tecnologia vestível, saúde e interação homem-máquina. Smartphones, em particular, destacam-se como dispositivos acessíveis e amplamente utilizados, equipados com sensores como acelerômetros e giroscópios capazes de coletar dados em tempo real sobre o movimento humano. Esses dados são cruciais para aplicações como monitoramento de saúde, treinamento esportivo e controle de dispositivos inteligentes.

Neste contexto, o UCI HAR Dataset representa uma base de dados amplamente utilizada para experimentos no campo de HAR. Ele foi construído a partir de experimentos realizados com 30 voluntários (19 a 48 anos), que executaram seis atividades distintas (walking, walking upstairs, walking downstairs, sitting, standing e laying). Os dados foram capturados por um smartphone posicionado na cintura dos participantes, registrando aceleração e velocidade angular em três eixos a uma taxa de 50Hz. Os sinais foram pré-processados com filtros para remover ruídos e segmentados em janelas fixas de 2,56 segundos com sobreposição de 50%. Dessa segmentação, extraíram-se 561 variáveis no domínio do tempo e da frequência, compondo o conjunto de features utilizado neste estudo.

O reconhecimento de padrões em dados complexos como os do UCI HAR Dataset é desafiador, especialmente em cenários onde rótulos não estão disponíveis ou são escassos. Nesse contexto, técnicas de aprendizado não supervisionado, como o algoritmo K-Means, oferecem um caminho promissor para identificar agrupamentos naturais nos dados. O presente trabalho tem como objetivo explorar a aplicação do K-Means para analisar os dados do UCI HAR Dataset, investigando sua capacidade de agrupar amostras com base em similaridades das features, sem supervisão explícita.

A relevância deste estudo reside na demonstração de que algoritmos não supervisionados podem servir como etapa preliminar em pipelines de HAR, auxiliando no entendimento inicial dos dados e identificando padrões úteis para algoritmos supervisionados. A metodologia empregada inclui o pré-processamento e normalização dos dados, redução de dimensionalidade com PCA e aplicação do K-Means, seguido de análise dos resultados com métricas como Silhouette Score e índice Rand ajustado.

Este artigo está estruturado da seguinte forma: a próxima seção detalha a metodologia empregada, incluindo descrição dos dados, técnicas de pré-processamento e configurações do algoritmo K-Means. Em seguida, apresenta-se a análise e discussão dos resultados, incluindo visualizações e métricas de qualidade dos agrupamentos. Por fim, são discutidas as principais conclusões e contribuições do estudo, bem como possibilidades de trabalhos futuros.

2. Metodologia

A metodologia deste estudo seguiu as etapas de pré-processamento, aplicação do algoritmo *K-Means* e avaliação dos resultados, conforme descrito a seguir:

1. **Pré-processamento dos Dados**
   O conjunto de dados foi carregado a partir dos arquivos disponibilizados no **UCI HAR Dataset**. Inicialmente, os dados de treinamento e teste foram separados, sendo utilizados apenas os arquivos *X_train.txt* e *X_test.txt*, que contêm as 561 *features* extraídas dos sinais de acelerômetro e giroscópio. Para normalizar os dados e garantir que as variáveis com diferentes escalas não dominassem o modelo, foi aplicada a técnica de normalização utilizando o *StandardScaler* da biblioteca **scikit-learn**.

2. **Redução de Dimensionalidade**
   Como o conjunto de dados possui um grande número de *features*, foi aplicada a análise de componentes principais (PCA) para reduzir a dimensionalidade, mantendo a maior parte da variância dos dados. O número de componentes foi ajustado para 2 para facilitar a visualização e interpretação dos resultados.

3. **Aplicação do Algoritmo K-Means**
   O algoritmo *K-Means* foi utilizado para realizar o agrupamento dos dados. O número de clusters (k) foi inicialmente determinado por meio do *método do cotovelo*, analisando a inércia para diferentes valores de k. Para a execução do *K-Means*, utilizou-se a implementação disponível na biblioteca **scikit-learn**, com k=6 clusters, considerando que existem 6 atividades distintas no conjunto de dados. O modelo foi treinado com os dados normalizados e os centros dos clusters foram identificados.

4. **Avaliação dos Resultados**
   Para avaliar a qualidade dos agrupamentos, foram utilizadas métricas como o *Silhouette Score*, que mede a coesão e separação dos clusters, e o índice Rand ajustado, que compara os agrupamentos obtidos com os rótulos reais das atividades. Além disso, a correspondência entre os clusters gerados e as atividades reais foi analisada visualmente por meio de gráficos bidimensionais, utilizando a redução de dimensionalidade do PCA.

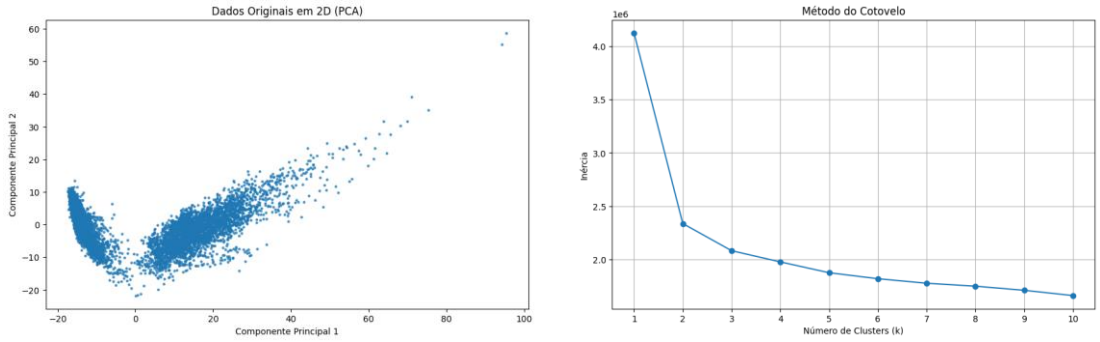
Esta metodologia é baseada em técnicas amplamente utilizadas em aprendizado de máquina não supervisionado, com adaptações para lidar especificamente com o conjunto de dados **UCI HAR Dataset**, garantindo uma análise eficaz dos padrões de movimento registrados pelos sensores.

3. Resultados e Discussão

O gráfico 1a apresenta os dados inicias para comparação. Após a aplicação do algoritmo *K-Means* no **UCI HAR Dataset**, o número ideal de clusters foi identificado utilizando o método do cotovelo (gráfico 1b). A análise da inércia para diferentes valores de k revelou que o ponto de inflexão ocorre em k=3 ou k=4, que não corresponde ao número de atividades presentes no conjunto de dados. Esse resultado sugere que o algoritmo *K-Means* pode não ser capaz de identificar agrupamentos corretos e

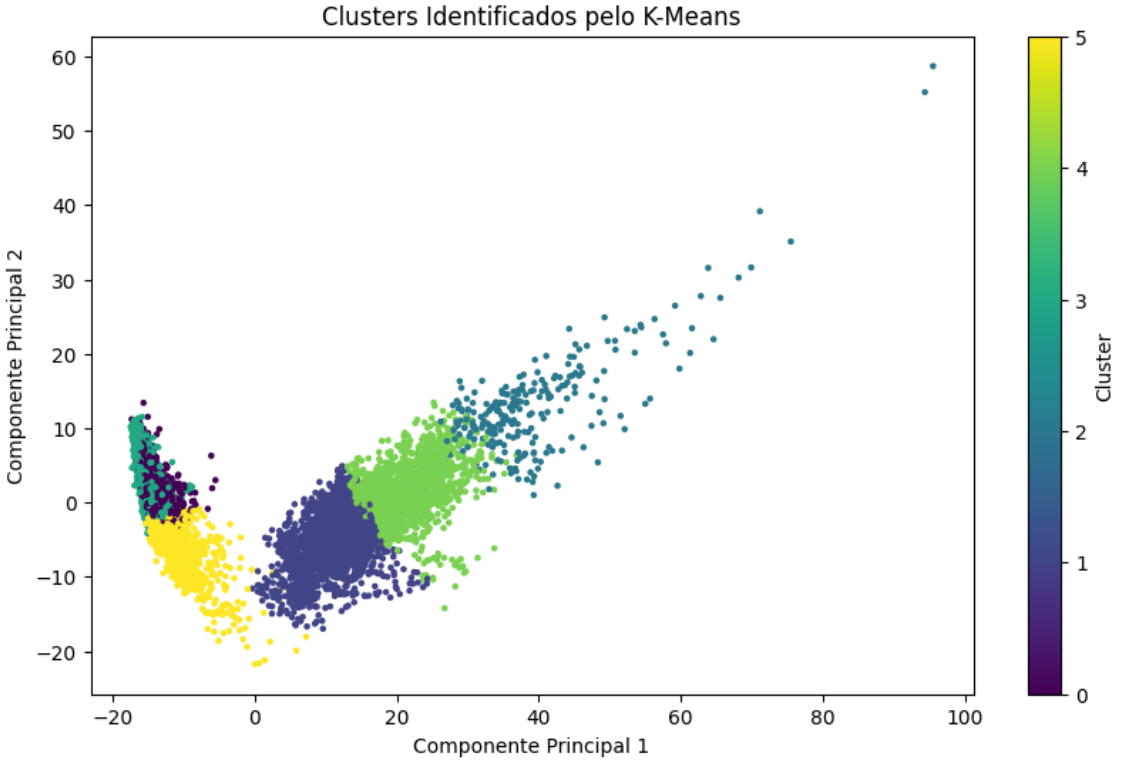
significativos para o conjunto de dados. Mesmo assim, o número de clusters foi mantido em k=6 como apresentado na descrição do dataset.

Gráfico 1: Método do cotovelo



A visualização dos clusters gerados pelo *K-Means* foi realizada utilizando a redução de dimensionalidade via PCA, o que permitiu a projeção dos dados em um espaço bidimensional. A partir dessa visualização, foi possível observar que os clusters estavam bem definidos, com as amostras agrupadas de maneira distinta.

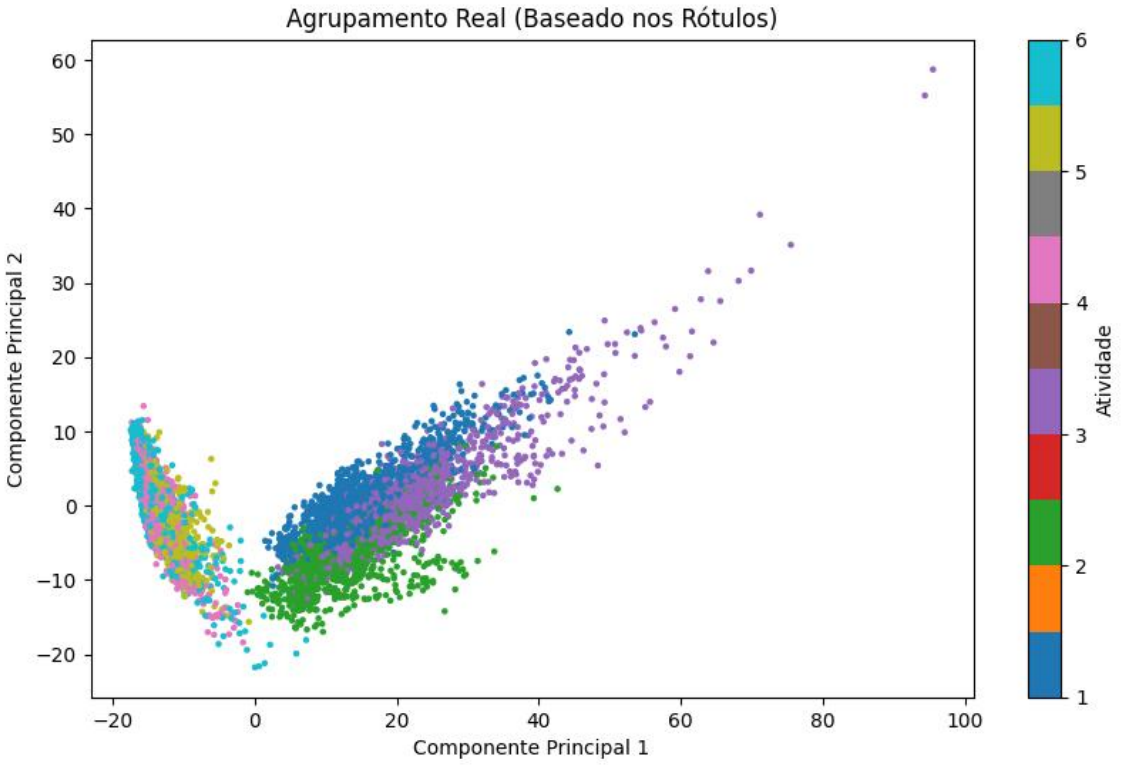
Gráfico 2: Gráfico dos Clusters Identificados



A qualidade do agrupamento foi avaliada através do cálculo do Silhouette Score, que obteve um valor de 0.1086410631864108. Esse valor indica uma baixa coesão interna entre os clusters e uma separação insatisfatória, sugerindo que o K-Means não conseguiu distinguir de forma clara as atividades realizadas. Além disso, o índice Rand ajustado foi calculado para comparar os clusters gerados com os rótulos reais das atividades, resultando em um valor de 0.41971368603551157. Este resultado reflete uma correspondência limitada entre os agrupamentos e os rótulos reais, apontando que o

algoritmo teve dificuldade em capturar corretamente as atividades realizadas pelos participantes.

Gráfico 3: Gráfico de Comparação com os Rótulos Reais



A comparação entre os clusters gerados e os rótulos reais revelou baixa correspondência geral, mesmo que atividades como *walking* e *walking upstairs* apresentem certa distinção. O gráfico mostra uma distribuição complexa dos dados reais, sem seguir o padrão de diagramas de Voronoi, evidenciando sobreposição significativa entre as atividades. Isso reforça que o K-Means tem limitações para lidar com esse tipo de problema, dada a ausência de fronteiras claras entre os clusters.

4. Conclusão e Trabalhos Futuros

Concluindo, o presente estudo demonstrou as dificuldades enfrentadas pelo algoritmo K-Means ao lidar com o conjunto de dados UCI HAR Dataset, um problema intrinsecamente desafiador devido à complexidade dos padrões de movimento e à alta dimensionalidade dos dados. Apesar do uso de técnicas de pré-processamento, redução de dimensionalidade e visualização, os resultados obtidos pelas métricas de avaliação, como o Silhouette Score (0.1086) e o índice Rand ajustado (0.4197), indicam uma correspondência limitada entre os clusters gerados e os rótulos reais das atividades.

Embora o método do cotovelo tenha sugerido k=3 ou k=4 como número ideal de clusters, optou-se por k=6 para respeitar o número de atividades no dataset. Ainda assim, os agrupamentos gerados não apresentaram separações claras, refletindo as limitações do K-Means em problemas com classes sobrepostas e padrões não lineares. A análise visual com PCA reforçou essa observação, evidenciando a sobreposição significativa entre os clusters, especialmente em atividades com características similares.

Esses resultados destacam a importância de explorar alternativas ao K-Means, como algoritmos que lidem melhor com fronteiras complexas e relações não lineares, a exemplo de DBSCAN ou métodos baseados em redes neurais. Para trabalhos futuros, também se sugere investigar técnicas mais avançadas de extração e seleção de features, visando melhorar a representatividade dos dados e facilitar a separação entre as classes. Assim, busca-se superar as limitações observadas e avançar na compreensão e modelagem de padrões de movimento humano.

5. Referências

SCIKIT-LEARN. scikit-learn: machine learning in Python. Disponível em: <https://scikit-learn.org/stable/>.

WIKIPEDIA CONTRIBUTORS. k-means clustering. Disponível em: <https://en.wikipedia.org/wiki/K-means_clustering>.

REYES-ORTIZ, J.; ANGUITA, D.; GHIO, A.; ONETO, L.; PARRA, X. **Human activity recognition using smartphones** [Dataset]. UCI Machine Learning Repository, 2013. Disponível em: https://doi.org/10.24432/C54S4K. Acesso em: 1 dez. 2024.