"IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE K-MEANS COM O DATASET HUMAN ACTIVITY RECOGNITION”

ISADORA COSTA MARQUES

ISRAEL ARAUJO FELIX DA SILVA

28 de novembro de 2024

Resumo

Este projeto apresenta o uso do algoritmo K-means, uma abordagem de aprendizado não supervisionado, para o reconhecimento de atividades humanas. Os dados foram coletados por sensores de aceleração e giroscópio, oferecendo informações detalhadas sobre os movimentos corporais.
O trabalho envolveu diversas etapas, incluindo análise exploratória, normalização dos dados, redução de dimensionalidade por PCA e definição do número ideal de clusters. Com uma pontuação média de Silhouette de 0,4768, o modelo conseguiu identificar padrões significativos, demonstrando uma separação razoável entre os clusters formados. Este relatório descreve o processo seguido, os resultados obtidos e as limitações enfrentadas.

Introdução

O reconhecimento de atividades humanas tem um papel cada vez mais relevante em áreas como monitoramento de saúde, dispositivos vestíveis e análise esportiva. A possibilidade de identificar atividades a partir de dados captados por sensores é essencial para criar soluções mais precisas e personalizadas.
Neste contexto, o algoritmo K-means foi escolhido por sua eficiência e simplicidade ao lidar com grandes volumes de dados. Ele permite agrupar amostras com base em semelhanças, auxiliando na identificação de padrões em conjuntos não rotulados.
O dataset utilizado, conhecido como Human Activity Recognition (HAR), contém 561 variáveis derivadas de medições de sensores de aceleração e giroscópio. Este relatório detalha como esses dados foram processados e como o K-means foi aplicado para identificar atividades humanas de maneira estruturada.

Metodologia

Etapa 1: Análise Exploratória

- **Carregamento e Visualização dos Dados:** Inicialmente, os dados foram carregados e inspecionados para identificar problemas como valores ausentes ou anomalias.
- **Matriz de Correlação:** Uma matriz de correlação foi criada para as 10 primeiras variáveis, permitindo explorar a relação entre elas e identificar redundâncias que poderiam impactar o agrupamento.

Etapa 2: Redução de Dimensionalidade com PCA

- Para simplificar a análise, aplicamos a Análise de Componentes Principais (PCA), que reduziu a dimensionalidade dos dados. As duas primeiras componentes explicaram 57,35% da variância, proporcionando uma visualização mais clara dos padrões.

Etapa 3: Normalização dos Dados

- **Motivação:** Dados de sensores têm escalas variadas, como aceleração (em metros por segundo) e giroscópio (em radianos por segundo).
- **Processo:** Padronizamos as variáveis usando o método StandardScaler, garantindo que todas tivessem média 0 e desvio padrão 1, equilibrando sua contribuição no agrupamento.

Etapa 4: Implementação do K-means

- **Inicialização:** Utilizamos o método K-means++, que posiciona os centróides de forma eficiente, reduzindo o risco de convergência para mínimos locais.
- **Escolha do Número de Clusters:** A análise pelo método do cotovelo indicou que 6 clusters eram ideais. Essa escolha foi validada pela pontuação de Silhouette, que confirmou uma separação razoável entre os grupos.

Etapa 5: Validação e Estabilidade

- Realizamos múltiplas execuções do K-means, com diferentes seeds, para garantir a consistência dos clusters formados.
- **Métricas de Avaliação:**
    - **Silhouette Score:** Pontuação média de 0,4768, refletindo uma separação razoável entre os clusters.
    - **Inércia:** Indicou uma boa coesão interna dos clusters.

Resultados

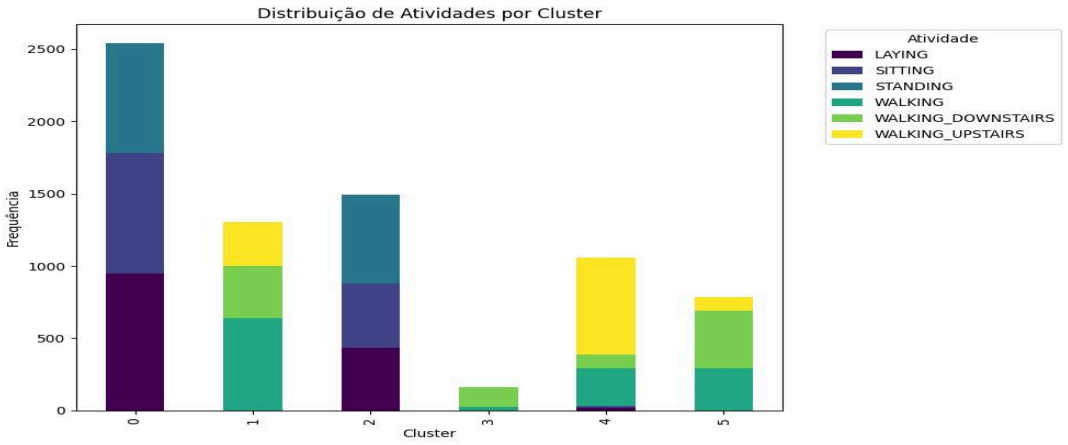
Métricas de Avaliação

- A pontuação de Silhouette indicou que os clusters estavam bem separados, embora com algumas sobreposições.

- A inércia demonstrou uma redução significativa até k=6, confirmando a escolha do número de clusters.

Visualizações

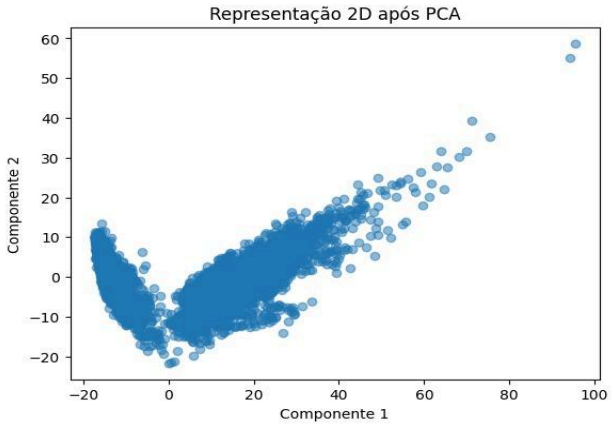
- **Gráfico PCA:** Exibiu os clusters em 2D, com centróides claramente visíveis, facilitando a interpretação visual.
- **Distribuição de Atividades:** Um gráfico de barras mostrou as atividades predominantes em cada cluster:
    - **Cluster 0:** Predominância de atividades estáticas, como deitar e sentar.
    - **Cluster 1:** Movimentos dinâmicos, como caminhar e descer escadas.
    - **Cluster 4:** Atividades como subir escadas.

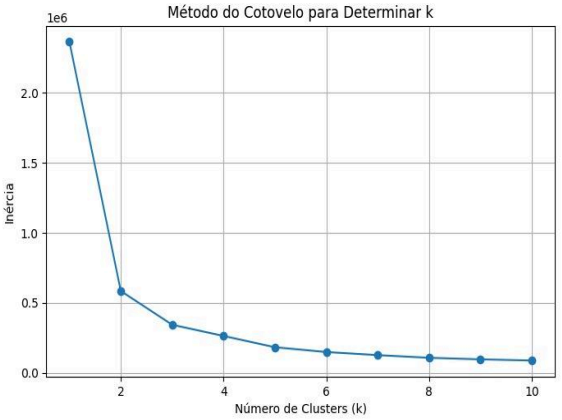


Discussão

Impacto das Escolhas do Modelo

- A redução de dimensionalidade com PCA ajudou na visualização, mas pode ter levado à perda de algumas informações relevantes.
- A escolha de 6 clusters foi adequada, pois os grupos capturaram atividades semelhantes de maneira consistente.





Limitações

- Algumas atividades, como ficar em pé e sentar, apresentaram sobreposição devido à similaridade entre as variáveis medidas.
- O K-means assume que os clusters têm forma esférica e tamanho similar, o que pode não refletir a complexidade dos dados reais.

Sugestões de Melhorias

- Testar algoritmos como DBSCAN, que não exigem um número fixo de clusters e lidam melhor com formatos variados.
- Explorar técnicas não lineares, como t-SNE, para identificar padrões mais complexos nos dados.

Conclusão

Este projeto demonstrou que o algoritmo K-means pode ser uma ferramenta eficaz para identificar atividades humanas a partir de dados de sensores. Apesar de suas limitações, o modelo conseguiu capturar padrões importantes, destacando o potencial do aprendizado não supervisionado na análise de dados complexos.

Trabalhos Futuros

- Incorporar técnicas supervisionadas para validar e refinar os clusters.
- Explorar algoritmos alternativos, como Gaussian Mixture Models ou DBSCAN, para lidar com limitações do K-means.
- Investir em estratégias de engenharia de variáveis para melhorar a separação entre clusters.