Relatório Técnico:

Residentes: Kira dos Reis Pinto; Lucas Gabriel Alcantara Silva

Data de entrega: 03/12/2024

1. RESUMO

Este trabalho apresenta a aplicação do algoritmo K-means para agrupar atividades humanas com base no dataset *Human Activity Recognition Using Smartphones*. Foram realizadas etapas de análise exploratória dos dados, normalização, redução de dimensionalidade e escolha do número ideal de clusters. O processo incluiu métricas de avaliação, como o Silhouette Score, para garantir a qualidade dos agrupamentos. Este relatório detalha as abordagens utilizadas e os resultados obtidos, destacando padrões relevantes no conjunto de dados analisado.

2. INTRODUÇÃO

O avanço da tecnologia de sensores embutidos em dispositivos móveis abriu novos horizontes para o reconhecimento de atividades humanas (HAR, na sigla em inglês). O dataset *Human Activity Recognition Using Smartphones*, disponibilizado pelo repositório UCI Machine Learning, oferece uma oportunidade única para explorar como sensores, como acelerômetros e giroscópios, podem ser usados para identificar padrões de movimento e atividades humanas. Este dataset inclui medições realizadas por 30 voluntários durante a realização de atividades diárias, como caminhar, subir escadas e ficar em pé, representando um conjunto de dados rico e desafiador para análise.

O objetivo deste trabalho é aplicar o algoritmo de clustering K-means para agrupar essas atividades, permitindo identificar padrões de comportamento humano a partir dos dados coletados pelos sensores. O relatório detalha todas as etapas do projeto, desde a análise exploratória dos dados até a avaliação dos resultados, utilizando métodos estatísticos e técnicas de visualização. Essa abordagem não apenas permite validar a eficácia do K-means neste contexto, mas também fornece insights sobre a estrutura dos dados e a interpretação dos clusters formados.

3. METODOLOGIA

3.1. Acesso ao Dataset

O conjunto de dados foi acessado diretamente do repositório UCI Machine Learning. Os arquivos principais utilizados foram:

- **X_train.txt**: Contendo os dados de entrada para o treinamento, com 561 variáveis calculadas a partir dos sinais brutos dos sensores.
- **y_train.txt**: Fornecendo os rótulos das atividades associadas aos dados de entrada, utilizados como referência para avaliação dos clusters formados.

Os dados foram carregados no ambiente de desenvolvimento e organizados em estruturas apropriadas utilizando bibliotecas Python como pandas e numpy.

3.2. Pré-processamento dos Dados

O pré-processamento dos dados foi uma etapa essencial para garantir a qualidade dos resultados. As seguintes ações foram realizadas:

- **Normalização**: Para padronizar a escala das variáveis, foi utilizada a classe StandardScaler da biblioteca scikit-learn. Esta etapa assegura que todas as variáveis contribuam de forma equilibrada para o cálculo das distâncias no algoritmo K-means.
- **Análise Exploratória**: Estatísticas descritivas foram geradas para identificar padrões e verificar a integridade dos dados. Esta etapa incluiu a verificação de valores ausentes e a análise de distribuições.

3.3. Redução de Dimensionalidade

Dada a alta dimensionalidade dos dados (561 variáveis), foi utilizada a técnica de Análise de Componentes Principais (PCA). Os dois primeiros componentes principais foram selecionados para:

- Facilitar a visualização dos clusters em um espaço bidimensional.
- Preservar a maior parte da variância dos dados originais (a variância explicada pelos dois componentes foi de aproximadamente 80%).

3.4. Implementação do Algoritmo K-means

O algoritmo K-means foi implementado utilizando a biblioteca scikit-learn. Os passos incluem:

I. **Escolha do número de clusters (K)**:
   A. Foi utilizado o método do cotovelo para determinar o valor ideal de K, baseado no gráfico de inércia.
   B. O valor de K foi ajustado para 6, com base na interpretação do gráfico e na coerência esperada dos clusters com as atividades do dataset.

II. **Treinamento do modelo**:
   A. O K-means foi configurado para utilizar a inicialização k-means++, melhorando a eficiência do algoritmo e reduzindo o risco de convergência para mínimos locais.

III. **Avaliação**:
   A. A métrica de silhouette score foi calculada para avaliar a qualidade dos clusters formados. Um valor médio de silhouette próximo de 0,67 indica uma boa separação entre os clusters.

3.5. Visualização dos Resultados

Para interpretar os resultados, as seguintes visualizações foram geradas:

- **Projeção PCA**: Um gráfico de dispersão mostrou os clusters no espaço bidimensional formado pelos dois primeiros componentes principais.
- **Distribuição dos clusters**: Usando o mapeamento dos clusters para as atividades originais, foi possível identificar a atividade predominante em cada cluster.

3.6. Documentação e Relatório

Todas as etapas foram documentadas de forma detalhada, e os resultados foram consolidados em um relatório técnico contendo gráficos, métricas de avaliação e uma discussão crítica sobre o desempenho do algoritmo e as limitações observadas. O código-fonte foi disponibilizado em um repositório no GitHub para facilitar a reprodução e o acompanhamento do projeto.

4.RESULTADOS

Os resultados obtidos a partir da aplicação do K-means destacaram a capacidade do algoritmo de identificar padrões significativos no dataset Human Activity Recognition Using Smartphones.

4.1. Métricas de Avaliação

Silhouette Score: O valor médio de 0,67 sugere uma separação sólida entre os clusters, com coesão interna suficiente para distinguir atividades específicas.
Inércia: A estabilização da inércia em
K=6
K=6 confirmou a adequação desse número de clusters, com uma redução significativa antes desse ponto e pouca variação posteriormente.

4.2. Visualizações e Análises

Projeção PCA:

Um gráfico de dispersão mostrou uma separação clara entre a maioria dos clusters, evidenciando o impacto da PCA em preservar a variância dos dados e facilitar a interpretação visual.
A maioria dos clusters apresentou distribuição bem definida, alinhada às atividades originais.

Distribuição dos Clusters:

Um histograma mapeando os clusters para as atividades mostrou uma predominância clara em atividades como "caminhar" e "subir escadas".
Atividades com padrões similares, como "sentado" e "em pé", apresentaram maior sobreposição, refletindo a dificuldade do algoritmo em diferenciá-las.
Gráfico de Silhouette:

A distribuição dos scores de silhouette por cluster destacou a qualidade geral do agrupamento, com valores positivos predominando e indicando clusters bem definidos.

5.DISCUSSÃO

A aplicação do algoritmo K-means ao dataset *Human Activity Recognition Using Smartphones* revelou tanto os pontos fortes quanto as limitações dessa abordagem no reconhecimento de atividades humanas.

Pontos Fortes:

- **Identificação de padrões:** O algoritmo conseguiu agrupar atividades distintas, como caminhar e subir escadas, de maneira eficaz.
- **Qualidade dos agrupamentos:** Um Silhouette Score de 0,67 indicou boa separação entre os clusters, validando o pré-processamento e a redução dimensional.
- **Visualização eficiente:** O uso da PCA facilitou a análise visual dos clusters, preservando aproximadamente 80% da variância.

Limitações:

- **Simplicidade do algoritmo:** O K-means pressupõe clusters esféricos e de tamanhos similares, o que pode não refletir a natureza dos dados.
- **Perda de informação:** A redução de dimensionalidade pode ter comprometido detalhes relevantes para a distinção de atividades semelhantes, como "em pé" e "sentado".
- **Sensibilidade à inicialização:** Apesar do uso de k-means++, a variabilidade dos resultados ainda depende da distribuição inicial dos clusters.
- **Falta de análise temporal:** A ausência de informações sobre transições entre atividades limitou a análise de comportamentos dinâmicos.

Discussão Geral:

Os resultados demonstram que o K-means é viável para identificar padrões em dados de sensores, mas não captura adequadamente atividades com características semelhantes. Estratégias futuras podem incluir algoritmos mais robustos, como DBSCAN ou Gaussian Mixture Models, e a incorporação de análises temporais para lidar com transições entre atividades.

6.CONCLUSÃO E TRABALHOS FUTUROS

A aplicação do algoritmo K-means no reconhecimento de atividades humanas revelou resultados promissores ao identificar padrões gerais nos dados. Atividades distintas, como caminhar e subir escadas, foram bem capturadas, evidenciando a eficácia do modelo em contextos de maior separação. A avaliação com o Silhouette Score reafirmou a qualidade dos agrupamentos, destacando o impacto positivo do pré-processamento e da redução dimensional. No entanto, o modelo demonstrou limitações em capturar nuances de atividades similares, como "em pé" e "sentado", e em lidar com clusters não esféricos ou de tamanhos desiguais. Essas restrições ressaltam a necessidade de estratégias mais avançadas para cenários mais complexos.

Trabalhos Futuros:

Exploração de Algoritmos Alternativos: Avaliar técnicas mais robustas, como DBSCAN ou Gaussian Mixture Models, para lidar com clusters de formatos variados e dados complexos.

Incorporação de Análise Temporal: Adicionar informações sequenciais para explorar transições entre atividades, o que pode melhorar a identificação de padrões dinâmicos.

Aprimoramento da Redução Dimensional: Experimentar técnicas como t-SNE ou UMAP para preservar mais informações relevantes ao reduzir a dimensionalidade dos dados.

Integração com Modelos Supervisionados: Combinar o clustering com modelos supervisionados para melhorar a interpretação e validação dos resultados.

Validação com Novos Datasets: Testar o modelo em outros conjuntos de dados para avaliar sua generalização em diferentes contextos de sensores e atividades.

Essas melhorias podem ampliar o alcance da abordagem e oferecer insights mais profundos para aplicações reais de reconhecimento de atividades humanas.

7.REFERÊNCIAS

BRAGANÇA, Hendrio Luis de Souza. *Reconhecimento de atividades humanas usando medidas estatísticas dos sensores inerciais dos smartphones*. 2019. 138 f. Dissertação (Mestrado em Informática) – Universidade Federal do Amazonas, Instituto de Computação, Manaus, 2019. Disponível em: https://tede.ufam.edu.br/handle/tede/7126. Acesso em: 3 dez. 2024..

UCI MACHINE LEARNING REPOSITORY. *Human Activity Recognition Using Smartphones Dataset*. Disponível em: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Acesso em: 3 dez. 2024.

FELCAM, Igor. *Entendendo Clusters e K-Means***.** CWI Software, 31 ago. 2020. Disponível em: https://medium.com/cwi-software/entendendo-clusters-e-k-means-56b79352b452. Acesso em: 3 dez. 2024.

SILVA, Renato. *Data Mining na Prática: Algoritmo K-Means*. DevMedia, 2012. Disponível em: https://www.devmedia.com.br/data-mining-na-pratica-algoritmo-k-means/4584. Acesso em: 3 dez. 2024.