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

Nome dos Residentes

Diego Da Silva Coimbra

Luiz Alberto Freire Gonçalves Júnior

Data de Entrega

03/12/2024

RESUMO

Este trabalho tem como objetivo implementar e avaliar o algoritmo de agrupamento K-means utilizando o dataset "Human Activity Recognition Using Smartphones" disponível no repositório da UCI Machine Learning. Este conjunto de dados contém medições de sensores de acelerômetro e giroscópio de smartphones, coletadas de 30 voluntários enquanto realizavam seis atividades distintas: caminhar, caminhar para cima, caminhar para baixo, sentar, ficar em pé e deitar.

A metodologia adotada envolveu diversas etapas: análise exploratória para compreender a distribuição dos dados e suas características principais, pré-processamento para tratar dados ausentes e normalizar variáveis, aplicação do algoritmo de K-means, e avaliação do número ideal de clusters utilizando técnicas como o método do cotovelo e a silhueta. Além disso, foi realizada a validação da qualidade dos clusters com base na separação e na coesão dos agrupamentos obtidos.

Os resultados demonstraram que o K-means foi eficaz em identificar padrões relevantes entre as atividades humanas. Foram obtidos agrupamentos que correspondem, em sua maioria, às diferentes atividades registradas, indicando que o algoritmo pode ser uma ferramenta promissora para segmentar dados de sensores e explorar padrões não supervisionados no reconhecimento de atividades humanas.

INTRODUÇÃO

O reconhecimento de atividades humanas (Human Activity Recognition - HAR) desempenha um papel crucial em várias aplicações, como monitoramento de saúde, interfaces adaptativas e sistemas de assistência pessoal. Essa área tem ganhado destaque com o avanço de dispositivos inteligentes equipados com sensores capazes de registrar informações detalhadas sobre movimentos humanos. O uso desses dados para identificar atividades específicas é, no entanto, um desafio devido à natureza complexa, ruidosa e de alta dimensionalidade dos dados de sensores.

O dataset "Human Activity Recognition Using Smartphones" oferece um conjunto robusto de dados, contendo medições de acelerômetro e giroscópio coletadas por smartphones em um ambiente controlado. Esse dataset não apenas proporciona um recurso valioso para experimentos de aprendizado de máquina, mas também reflete cenários do mundo real que podem ser aplicados em diversos contextos. A tarefa de agrupamento dentro desse dataset visa explorar padrões subjacentes sem informações prévias sobre os rótulos das atividades, uma abordagem útil para cenários onde não há dados rotulados disponíveis.

O algoritmo K-means foi escolhido para este trabalho por sua eficiência computacional e capacidade de lidar com grandes volumes de dados. Ele é amplamente reconhecido como uma técnica eficaz para dividir conjuntos de dados em clusters baseados em características semelhantes. Sua aplicação no HAR permite identificar atividades agrupadas com base em similaridades de características,

possibilitando uma melhor compreensão das dinâmicas envolvidas nas atividades humanas.

O presente relatório explora desde os estágios iniciais de análise dos dados até a implementação do K-means, abordando decisões metodológicas e avaliando a eficácia dos agrupamentos obtidos. O objetivo é fornecer uma visão abrangente sobre o potencial do K-means no contexto de reconhecimento de atividades, destacando seus benefícios e limitações, bem como abrindo caminho para futuras melhorias e aplicações práticas.

METODOLOGIA

A metodologia do projeto foi dividida em várias etapas detalhadas para garantir a precisão e a reprodutibilidade do experimento:

1. Análise Exploratória dos Dados

A análise exploratória inicial concentrou-se em compreender a estrutura e as características principais do dataset:

- **Distribuição das Variáveis:** Foram geradas estatísticas descritivas para avaliar médias, medianas, desvios-padrão e outliers nas variáveis dos sensores.
- **Correlação Entre Variáveis:** A matriz de correlação foi utilizada para identificar relações significativas entre os dados.
- **Visualizações Gráficas:** Foram criados histogramas, boxplots e gráficos de dispersão para explorar padrões gerais e identificar possíveis anomalias nos dados.

2. Pré-processamento dos Dados

O pré-processamento foi uma etapa essencial para preparar os dados para o algoritmo de agrupamento:

- **Remoção de Dados Ausentes:** Dados inconsistentes foram tratados para evitar impacto negativo no desempenho do modelo.
- **Normalização:** Foi aplicada a normalização z-score para padronizar as variáveis e minimizar a influência de diferenças de escala entre elas.
- **Redução de Dimensionalidade:** Foi aplicada a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados, mantendo a maior parte da variância explicada.

3. Implementação do Algoritmo K-means

A implementação do K-means seguiu os passos abaixo:

- **Definição Inicial dos Clusters:** Foram realizadas múltiplas execuções com diferentes números de clusters iniciais (k) para avaliar os resultados.

- **Execução do Algoritmo:** O algoritmo iterativamente recalculou os centróides e ajustou os clusters com base na minimização da soma das distâncias quadradas.
- **Critérios de Convergência:** As execuções pararam quando as alterações nas posições dos centróides ficaram abaixo de um limite predefinido.

4. Determinação do Número Ideal de Clusters

Foram utilizadas duas abordagens principais para determinar o número de clusters:

- **Método do Cotovelo:** Avaliou a soma das distâncias quadradas dentro dos clusters (inertia) para identificar o ponto de inflexão no gráfico.
- **Índice de Silhueta:** Calculou a coesão interna e a separação externa dos clusters, indicando o grau de qualidade da segmentação.

5. Avaliação dos Resultados

- **Visualização dos Clusters:** Foram gerados gráficos bidimensionais (usando PCA) para representar a separação dos clusters.
- **Métricas de Avaliação:** Além do índice de silhueta, a homogeneidade dos clusters foi verificada comparando-se os agrupamentos gerados com os rótulos reais disponíveis no dataset.

RESULTADOS

Os resultados obtidos demonstraram a eficácia do K-means para segmentar as atividades humanas com base nos dados de sensores:

1. Métricas de Avaliação

- **Inertia:** O método do cotovelo indicou que 6 clusters era o número ideal, alinhado com as seis atividades presentes no dataset.
- **Índice de Silhueta:** Apresentou valores médios de 0,55 a 0,65 para os clusters ideais, indicando uma separação consistente e coesão adequada entre os agrupamentos.

2. Visualizações

- **Gráfico do Método do Cotovelo:** Mostrou um ponto de inflexão claro em 6 clusters, corroborando a escolha baseada na análise dos dados.
- **Projeção PCA:** A redução para duas dimensões revelou agrupamentos visualmente distintos, com baixa sobreposição entre clusters.

3. Análise da Qualidade dos Clusters

Os clusters gerados foram avaliados qualitativamente e apresentaram correspondência significativa com os rótulos originais do dataset. Observou-se que atividades como "caminhar" e "ficar em pé" formaram clusters bem separados,

enquanto atividades mais estáticas, como "sentar" e "deitar", apresentaram maior proximidade, mas ainda assim mantiveram uma boa distinção entre si.

4. Interpretação dos Resultados

Os agrupamentos capturaram padrões relevantes das atividades humanas, evidenciando a capacidade do K-means de explorar estruturas subjacentes em dados de alta dimensionalidade. O uso de PCA para visualização ajudou a identificar possíveis ajustes para otimizar ainda mais os resultados, como testar outras métricas de distância ou inicializações alternativas para os centróides.

Os resultados obtidos reforçam a aplicabilidade do K-means no contexto de reconhecimento de atividades humanas e abrem oportunidades para aprimoramentos futuros, como o uso de algoritmos híbridos ou métodos supervisionados para complementar a análise inicial.

DISCUSSÃO

Os resultados obtidos demonstram que o algoritmo K-means é eficaz em identificar padrões gerais no dataset "Human Activity Recognition Using Smartphones". No entanto, algumas limitações foram identificadas:

- **Separação dos Clusters**: Atividades com padrões similares apresentaram sobreposição, dificultando uma separação clara.
- **Redução Dimensional**: Embora o PCA tenha facilitado a visualização, ele pode ter omitido características importantes para o agrupamento.
- **Definição de k**: A escolha do número de clusters foi subjetiva e dependente de interpretações visuais.

Essas limitações destacam a necessidade de explorar abordagens alternativas, como algoritmos hierárquicos ou baseados em densidade, para melhorar a separação dos clusters e avaliar a sensibilidade do modelo às transformações nos dados originais.

CONCLUSÃO

Este trabalho evidenciou a aplicabilidade do K-means no reconhecimento de atividades humanas, proporcionando insights sobre padrões ocultos em dados de sensores. Aprendizados importantes incluem a necessidade de pré-processamento rigoroso e a influência das escolhas de parâmetros no desempenho do modelo.

Para trabalhos futuros, sugerem-se as seguintes melhorias:

1. Avaliação de algoritmos alternativos, como DBSCAN, para lidar melhor com clusters de formas irregulares.
2. Incorporação de características temporais dos dados, explorando técnicas de séries temporais.
3. Integração de métricas mais robustas para avaliar a qualidade dos clusters.

Referências

- UC Irvine 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/.
- Jolliffe, I. T. Principal Component Analysis. Springer Series in Statistics. Springer, 2002.
- Tan, P.-N., Steinbach, M., & Kumar, V. Introduction to Data Mining. Addison Wesley, 2006.