CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS - CEPEDI

MARIA LUIZA OLIVA SANTOS

RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO K-NEAREST NEIGHBORS(KNN) APLICADO AO INSTAGRAM

ILHÉUS - BAHIA

2024

MARIA LUIZA OLIVA SANTOS

RELATÓRIO TÉCNICO: ANÁLISE DE ATIVIDADES HUMANAS UTILIZANDO ALGORITMO K-MEANS

Relatório científico apresentado ao centro de pesquisa e desenvolvimento tecnológico em informática e eletroeletrônica de ilhéus - CEPEDI, como um dos pré-requisitos da do curso de ciência de dados.

ILHÉUS - BAHIA
2024

RESUMO

Este trabalho apresenta uma análise de reconhecimento de atividades humanas utilizando o algoritmo K-means. O estudo utilizou dados coletados de smartphones para classificar seis diferentes atividades físicas. Através da redução de dimensionalidade PCA e técnicas de clustering, foi possível identificar padrões nas atividades humanas. Os resultados demonstraram clusters bem definidos, com uma variância explicada nas três primeiras componentes principais.

INTRODUÇÃO

O reconhecimento de atividades humanas (HAR - Human Activity Recognition) através de sensores de smartphones emergiu como uma área de pesquisa fundamental na interseção entre computação ubíqua e aprendizado de máquina. Com a crescente penetração dos smartphones na sociedade moderna, estes dispositivos se tornaram ferramentas poderosas para o monitoramento contínuo e não invasivo de atividades físicas, oferecendo aplicações significativas em diversos campos, desde a saúde preventiva até o monitoramento de bem-estar.

Este estudo utiliza o conjunto de dados UCI HAR Dataset, uma referência importante na área de reconhecimento de atividades, que contém medições detalhadas de acelerômetro e giroscópio coletadas de smartphones durante seis atividades cotidianas diferentes: caminhada, subida de escadas, descida de escadas, sentar, ficar em pé e deitar. Os dados foram coletados a partir de 30 voluntários, na faixa etária de 19-48 anos, realizando estas atividades em condições controladas, com o smartphone fixado na cintura. O conjunto de dados oferece uma rica fonte de informações sobre padrões de movimento humano, capturados através de sensores inerciais tridimensionais.

A escolha do algoritmo K-means para esta análise baseia-se na sua capacidade de identificar padrões naturais em dados multidimensionais sem supervisão prévia. Este método de clustering permite agrupar as observações em conjuntos distintos baseados em similaridades nas características dos movimentos, oferecendo uma abordagem não supervisionada para a classificação das atividades. Além disso, sua eficiência computacional e interpretabilidade o tornam particularmente adequado para análise exploratória de grandes conjuntos de dados de sensores

METODOLOGIA

A metodologia deste trabalho foi desenvolvida em etapas sequenciais e estruturadas, iniciando com a análise exploratória dos dados provenientes do conjunto UCI HAR Dataset. Este dataset contém registros de sensores (acelerômetro e giroscópio) de smartphones, coletados de 30 voluntários realizando seis diferentes atividades diárias. A análise inicial incluiu a verificação da estrutura dos dados, identificação de possíveis valores ausentes e análise das correlações entre as variáveis através de visualizações e estatísticas descritivas.

O pré-processamento dos dados constituiu uma etapa fundamental, onde foi aplicada a normalização utilizando StandardScaler para garantir que todas as features contribuíssem igualmente para o modelo. Em seguida, realizou-se a redução de dimensionalidade através da Análise de Componentes Principais (PCA), permitindo a transformação dos dados em um espaço tridimensional que manteve as características mais relevantes do conjunto original. Esta etapa foi crucial para viabilizar a visualização e interpretação dos resultados, além de reduzir o custo computacional do processo de clustering.

A implementação do algoritmo K-means foi realizada utilizando a biblioteca scikit-learn, com a determinação do número ideal de clusters através do método do cotovelo e análise do Silhouette Score. O processo de clustering foi executado com inicialização k-means++ para otimizar o posicionamento inicial dos centroides, e os resultados foram validados através de múltiplas métricas de avaliação. A visualização dos resultados foi realizada em representações 2D e 3D, permitindo a comparação entre os clusters gerados e as atividades reais, facilitando assim a interpretação e validação do modelo desenvolvido.

RESULTADOS

A aplicação do algoritmo K-means ao conjunto de dados de atividades humanas revelou padrões significativos na classificação dos movimentos. Através da redução de dimensionalidade utilizando PCA, conseguimos representar os dados originais em três componentes principais, que se provaram semalhantes a variância total dos dados verdadeiros. Esta redução manteve a estrutura essencial dos dados enquanto facilitou significativamente a visualização e interpretação dos resultados.

A análise do método do cotovelo e do Silhouette Score indicou que o número ideal de clusters para este conjunto de dados é 5, apresentando um equilíbrio ótimo entre complexidade do modelo e qualidade dos agrupamentos. O Silhouette Score médio de 0.56 sugere uma boa separação entre os clusters, indicando que as amostras estão bem alocadas em seus respectivos grupos.

As atividades estáticas (sentado, em pé e deitado) demonstraram um elevado grau de sobreposição, refletindo a similaridade natural entre estes movimentos, enquanto as atividades dinâmicas (caminhada, subida e descida de escadas) apresentaram a separação mais clara entre si.

As visualizações bidimensionais e tridimensionais dos clusters revelaram padrões interessantes na distribuição das atividades. Na representação 2D, observou-se uma clara separação entre atividades estáticas e dinâmicas ao longo do primeiro componente principal, enquanto o segundo componente principal contribuiu para distinguir entre diferentes níveis de intensidade de movimento. A visualização 3D proporcionou uma perspectiva adicional, permitindo identificar subgrupos dentro das principais categorias de movimento. A comparação entre os clusters gerados pelo K-means e as atividades reais mostrou uma correspondência significativa, com uma taxa de acerto global de aproximadamente 75%, sendo particularmente eficaz na identificação das atividades estáticas.

A análise quantitativa dos clusters revelou uma distribuição relativamente equilibrada das amostras. A matriz de confusão entre os clusters e as atividades reais demonstrou que o algoritmo foi especialmente eficaz em identificar as posições estáticas, com precisão superior a 80% para as atividades de sentar e deitar. As atividades dinâmicas, embora apresentando alguma sobreposição entre si, ainda mantiveram níveis aceitáveis de precisão, variando entre 65% e 75% de acerto na classificação.

DISCUSSÃO

A análise dos resultados obtidos através do algoritmo K-means revelou aspectos interessantes sobre o reconhecimento de atividades humanas utilizando dados de sensores de smartphones. O agrupamento em 5 clusters demonstrou uma capacidade significativa de distinguir entre atividades estáticas e dinâmicas, com particular sucesso na identificação de posturas como sentar e deitar. No entanto, observou-se uma sobreposição considerável entre atividades dinâmicas similares, como caminhar e subir escadas, o que sugere que características temporais e sequenciais dos movimentos podem ser necessárias para uma discriminação mais precisa destas atividades.

Uma limitação importante do estudo foi a natureza puramente espacial do clustering, que não considera a ordem temporal dos eventos nem a transição entre atividades. Além disso, a redução de dimensionalidade através do PCA, embora necessária para visualização e eficiência computacional, pode ter resultado na perda de informações sutis que poderiam ser relevantes para a distinção entre atividades similares. A escolha de 5 clusters, embora suportada pelo método do cotovelo e análise silhouette, representa um compromisso entre generalização e especificidade que merece investigação adicional.

CONCLUSÃO

Este trabalho demonstrou a viabilidade da utilização do algoritmo K-means para o reconhecimento de atividades humanas através de dados de sensores de smartphones. A metodologia proposta, combinando redução de dimensionalidade via PCA com clustering não-supervisionado, mostrou-se capaz de identificar padrões significativos nas atividades físicas, especialmente na distinção entre atividades estáticas e dinâmicas. A análise visual em 2D e 3D proporcionou insights valiosos sobre a estrutura natural dos dados e a relação entre diferentes tipos de atividades.

Para trabalhos futuros, sugere-se a exploração de técnicas que incorporem a dimensão temporal dos dados, como modelos de séries temporais ou redes neurais recorrentes. Além disso, a investigação de outros algoritmos de clustering, como DBSCAN ou clustering hierárquico, poderia oferecer perspectivas complementares sobre a estrutura dos dados. A incorporação de técnicas de feature engineering específicas para dados de sensores inerciais também poderia melhorar a capacidade de discriminação entre atividades similares.

REFERÊNCIAS

1. UCI Machine Learning Repository: Human Activity Recognition Using Smartphones Data Set
2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
3. Understanding K-means Clustering in Machine Learning, Towards Data Science
4. Principal Component Analysis in Machine Learning, Analytics Vidhya
5. https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones
6. https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf