Relatório Técnico: Reconhecimento de Atividades Humanas com K-means

Nome do Aluno: Fagner de Oliveira Carrena

Nome do Aluno: Sthefane dos Santos Ramos

Data de Entrega: 28/11/2024

Resumo

Este projeto tem como objetivo explorar técnicas de clustering para o reconhecimento de atividades humanas a partir do conjunto de dados UCI HAR (Human Activity Recognition). Utilizando o algoritmo K-means, foram analisadas as atividades físicas de um grupo de participantes a partir de medições de acelerômetros e giroscópios. O modelo foi implementado em Python, utilizando bibliotecas como pandas, scikit-learn e seaborn. A principal conclusão foi que o K-means foi capaz de identificar quatro clusters representando diferentes atividades físicas, mas com algumas limitações na separação dos dados. O índice de Silhouette para o modelo foi de 0.33, sugerindo uma qualidade moderada nos agrupamentos. A análise revelou que técnicas de clustering podem ser aplicadas de maneira eficaz para o reconhecimento de atividades humanas, mas melhorias no pré-processamento e na escolha do número de clusters são necessárias para otimizar os resultados.

Introdução

Contextualização do Problema:
O reconhecimento de atividades humanas é uma área de pesquisa importante dentro do campo da análise de dados, com aplicações em diversos setores, como saúde, esporte, segurança e inteligência artificial. O objetivo principal é identificar padrões de comportamento humano a partir de dados coletados por sensores, como acelerômetros e giroscópios, frequentemente usados em dispositivos móveis e wearables. O desafio está em interpretar essas grandes quantidades de dados e reconhecer automaticamente as atividades realizadas por um indivíduo, como caminhar, correr, subir escadas, entre outras.

Neste projeto, foi utilizado o conjunto de dados UCI HAR (Human Activity Recognition), que contém leituras de sensores de aceleração e giroscópio provenientes de smartphones, com o objetivo de classificar diferentes atividades físicas. As características do conjunto de dados, que incluem 561 variáveis para cada amostra de atividade, tornam o problema ainda mais desafiador, exigindo o uso de técnicas de redução de dimensionalidade e clustering para explorar e entender os padrões presentes nos dados.

O K-means foi escolhido como técnica de clustering devido à sua simplicidade, eficiência e capacidade de lidar com grandes volumes de dados, como os encontrados no conjunto UCI HAR. O K-means é um algoritmo de agrupamento não supervisionado que segmenta os dados em K clusters, de acordo com a proximidade de pontos no espaço de características. Dada a natureza dos dados, que envolvem medições contínuas

de diversas variáveis, o K-means é uma ferramenta poderosa para identificar agrupamentos naturais nos dados sem a necessidade de rótulos previamente definidos.

A escolha do K-means se justifica também pela sua capacidade de fornecer uma divisão clara dos dados, o que facilita a análise de como diferentes atividades podem ser agrupadas em classes distintas, ajudando na compreensão dos padrões de movimento humano.

Metodologia

A metodologia foi estruturada em três etapas principais: análise exploratória, implementação do K-means e avaliação dos clusters formados.

1. **Análise Exploratória dos Dados**:
   A primeira etapa do projeto envolveu a análise exploratória dos dados, com o objetivo de compreender as características do conjunto de dados e identificar possíveis padrões. Para isso, foi realizada uma análise descritiva das variáveis e calculadas as estatísticas principais (média, desvio padrão, mínimo, máximo e percentis). Também foi gerado um mapa de correlação para visualizar a relação entre as diferentes variáveis e verificar se existem correlações fortes entre as medições, o que pode indicar redundância ou agrupamentos naturais nos dados.
2. **Implementação do K-means**:
   A implementação do K-means foi realizada após a análise exploratória, com o objetivo de segmentar os dados em grupos significativos. Antes de aplicar o algoritmo de clustering, foi necessário normalizar os dados utilizando a técnica de **StandardScaler** para garantir que todas as variáveis tivessem a mesma escala e, assim, evitar que variáveis com maiores magnitudes dominassem o processo de agrupamento.
3. **Escolha do Número de Clusters**:
   Para determinar o número ideal de clusters, foi utilizado o **método do cotovelo**, que analisa a inércia (ou soma das distâncias quadradas dentro dos clusters) em função do número de clusters. A ideia é escolher o valor de K onde a inércia começa a se estabilizar. Além disso, o **índice de Silhouette** foi calculado para avaliar a coesão e separação dos clusters formados, sendo um indicador importante da qualidade do agrupamento. O número de clusters que maximiza o índice de Silhouette foi escolhido como o valor ideal.

Resultados

Métricas de Avaliação:
O número de clusters foi determinado como sendo K=4, com base nos resultados obtidos pelo método do cotovelo e pelo índice de Silhouette. O **índice de Silhouette** para K=4 foi de 0.33, indicando uma separação moderada entre os clusters, o que sugere que o modelo conseguiu formar agrupamentos razoavelmente distintos, mas com alguma sobreposição entre as classes.

Gráficos e Análise dos Clusters:
A seguir, apresentamos os resultados visuais do K-means aplicado com K=4, após a redução de dimensionalidade via PCA. A visualização dos dados em duas dimensões (componente principal 1 e componente principal 2) permite observar como os clusters foram formados no espaço de características.

- O gráfico de dispersão gerado mostra a distribuição dos dados ao longo dos dois primeiros componentes principais, coloridos de acordo com os clusters formados pelo algoritmo K-means. A separação entre os clusters é visível, embora haja alguma sobreposição, o que é esperado em problemas de clustering com dados complexos.
- O método do cotovelo revelou uma diminuição significativa na inércia à medida que o número de clusters aumentava de 1 para 4, mas sem uma grande melhoria após K=4, indicando que este é o número ideal de clusters.

Análise da Qualidade dos Clusters:
A análise visual e os valores das métricas indicam que os clusters formados são relativamente coesos e bem separados, o que sugere que o K-means conseguiu segmentar as atividades humanas de forma eficaz. No entanto, um Silhouette Score de 0.33 indica que há espaço para melhorias, o que pode ser atribuído à natureza complexa dos dados e à possível necessidade de técnicas mais avançadas de clustering ou de pré-processamento dos dados.

Discussão

Os resultados obtidos indicam que o K-means foi capaz de identificar quatro clusters distintos no conjunto de dados de reconhecimento de atividades humanas. A visualização dos clusters após a redução de dimensionalidade com PCA mostra uma separação razoável entre as atividades, com algumas áreas de sobreposição. O Silhouette Score de 0.33 sugere que, embora a segmentação tenha sido razoavelmente boa, a qualidade do agrupamento ainda pode ser aprimorada. Esse resultado está alinhado com a complexidade do conjunto de dados, que contém variáveis altamente correlacionadas e interdependentes.

Limitações do Modelo:
O principal desafio neste projeto foi a grande quantidade de variáveis no conjunto de dados (561 características), o que pode ter contribuído para a dificuldade de segmentação clara. Embora a técnica de PCA tenha sido utilizada para reduzir a dimensionalidade, algumas informações importantes podem ter sido perdidas, o que pode afetar a qualidade dos clusters. Além disso, o algoritmo K-means assume que os clusters são esféricos e de tamanho semelhante, o que nem sempre é o caso em dados reais. Portanto, pode ser que algumas atividades tenham sido mal agrupadas devido à forma dos dados.

Outra limitação do modelo foi a escolha do K-means, que é sensível à inicialização dos centros de clusters. Embora tenha sido utilizado o método 'k-means++' para melhorar a escolha inicial dos centroids, o algoritmo ainda pode ter convergido para um ponto

subótimo. Além disso, a normalização das variáveis foi um passo crucial, mas ela pode não ter sido suficiente para lidar com todas as nuances dos dados, como as diferenças nas unidades de medida entre os sensores.

Impacto das Escolhas Feitas:
A escolha do K-means foi fundamental para a análise, pois possibilitou a segmentação dos dados de forma rápida e eficiente. No entanto, a escolha de K=4 como o número ideal de clusters foi baseada em métodos heurísticos, como o método do cotovelo e o índice de Silhouette, o que pode não ter sido o valor mais preciso. Alternativas, como o uso de métodos de clustering hierárquico ou DBSCAN, poderiam fornecer uma segmentação mais robusta em casos com dados mais complexos ou com formas de clusters não esféricas.

A redução de dimensionalidade com PCA, embora útil para visualização, pode ter limitado a capacidade do modelo de capturar toda a variabilidade nos dados. Em cenários com mais variáveis e maior complexidade, técnicas de redução de dimensionalidade mais avançadas, como t-SNE ou UMAP, poderiam ser exploradas para melhorar a visualização e o desempenho do clustering.

Conclusão e Trabalhos Futuros

Aprendizados Principais:
Este projeto permitiu uma compreensão profunda sobre as técnicas de clustering e redução de dimensionalidade aplicadas a dados de reconhecimento de atividades humanas. A principal conclusão foi que o K-means, combinado com PCA, pode ser uma abordagem eficaz para segmentar dados complexos, mas que a qualidade da segmentação depende fortemente da escolha do número de clusters e da técnica de pré-processamento. A análise exploratória também revelou a importância de avaliar a correlação entre as variáveis antes de aplicar algoritmos de clustering.

Sugestões para Trabalhos Futuros:

1. **Aprimoramento do Pré-processamento**:
   Uma possível melhoria seria aplicar técnicas de seleção de características para reduzir ainda mais a dimensionalidade do conjunto de dados sem perder informações relevantes. Métodos como o uso de LASSO (Least Absolute Shrinkage and Selection Operator) ou modelos baseados em árvores de decisão podem ser úteis para selecionar variáveis mais relevantes.
2. **Exploração de Outros Algoritmos de Clustering**:
   Como o K-means pode não ser ideal para todos os tipos de dados, a exploração de outros métodos de clustering, como DBSCAN ou clustering hierárquico, pode resultar em melhores agrupamentos, especialmente quando os dados não formam clusters esféricos. A experimentação com esses algoritmos pode melhorar a coesão e a separação dos clusters.
3. **Ajuste de Parâmetros do Modelo**:
   O número de clusters foi escolhido com base em métodos heurísticos, mas uma análise mais rigorosa, como a validação cruzada ou a otimização de hiperparâmetros, poderia ser feita para garantir a escolha mais robusta de K.

4. **Exploração de Modelos Supervisionados**:
   Para uma análise mais aprofundada e precisa, seria interessante explorar modelos supervisionados de classificação, onde rótulos de atividades são utilizados para treinar modelos, como Support Vector Machines (SVM) ou Redes Neurais. Isso pode permitir uma abordagem mais acurada para o reconhecimento de atividades humanas.

Referências

1. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones. *21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.
2. Scikit-learn Documentation. (2024). K-means clustering. Recuperado de: https://scikit-learn.org/stable/modules/clustering.html#k-means
3. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
5. Seaborn Documentation. (2024). Seaborn: statistical data visualization. Recuperado de: https://seaborn.pydata.org/