Relatório Técnico: Análise de Agrupamento de Atividades Humanas com K-MEANS

Matheus Tamarindo Gonzaga

Rafael Ottoni Rodrigues Gonçalves

Data de Entrega: 03/12/2024

O objetivo deste projeto foi aplicar o algoritmo K-MEANS para agrupar dados de atividades humanas coletados por sensores. O conjunto de dados, composto por medições de 561 variáveis que representam sinais brutos dos sensores, é ideal para tarefas de análise de agrupamento. Para garantir a eficiência do modelo, o processo envolveu a normalização dos dados, a escolha do número ideal de clusters utilizando métodos como o cotovelo (*Elbow Method*) e o *Silhouette Score*, além da redução da dimensionalidade por meio do PCA.

O modelo foi avaliado usando métricas como inércia e *Silhouette Score*, com visualizações em 2D e 3D dos clusters para facilitar a interpretação dos resultados. Os resultados demonstraram uma boa separação entre os clusters, juntamente com uma coesão interna significativa.

INTRODUÇÃO

O Reconhecimento de Atividade Humana (HAR) é uma área crucial no campo do aprendizado de máquina, com múltiplas aplicações, tais como monitoramento da saúde, interação homem-máquina e automação. Este projeto teve como objetivo agrupar atividades humanas utilizando dados de sensores, onde cada amostra do conjunto de dados é composta por medições de variáveis sensoriais que capturam diferentes aspectos do comportamento humano.

O algoritmo K-MEANS foi selecionado para o agrupamento devido à sua simplicidade, eficiência e ampla utilização em tarefas similares. Um dos principais desafios deste projeto reside na alta dimensionalidade dos dados, o que torna necessário o emprego de técnicas como a Análise de Componentes Principais (PCA) para a redução dimensional. Adicionalmente, a escolha criteriosa do número de clusters é fundamental para assegurar que os agrupamentos representem de forma precisa as diferentes atividades.

SOBRE O CONJUNTO DE DADOS

O conjunto de dados HAR foi obtido do repositório de conjuntos de dados UCI[1]. Este conjunto de dados foi coletado de um grupo de 30 voluntários com idades variando de 19 a 48 anos selecionados para esta tarefa, realizando diferentes atividades com um smartphone na cintura. Os dados foram registrados com a ajuda de sensores (acelerômetro e giroscópio) presentes naquele smartphone.

Cada pessoa realizou seis atividades (ANDAR, ANDAR_NO ANDAR DE CIMA, ANDAR_NO ANDAR DE BAIXO, SENTAR, FICAR DE PÉ, DEITAR) usando um smartphone (Samsung Galaxy S II) na cintura. Usando seu acelerômetro e giroscópio incorporados, capturamos aceleração linear de 3 eixos e velocidade angular de 3 eixos a uma taxa constante de 50 Hz. Os experimentos foram gravados em vídeo para rotular os dados manualmente. O conjunto de dados obtido foi particionado aleatoriamente em dois conjuntos, onde 70% dos voluntários foram

[1] UC Irvine Machine Learning Repository

selecionados para gerar os dados de treinamento e 30% os dados de teste.

INFORMAÇÕES DE ATRIBUTO

Para cada registro no conjunto de dados, o seguinte é fornecido:

- Aceleração triaxial do acelerômetro (aceleração total) e a aceleração estimada do corpo.
- Velocidade angular triaxial do giroscópio.
- Um vetor de 561 características com variáveis de domínio de tempo e frequência.
- Seu rótulo de atividade.
- Um identificador do sujeito que realizou o experimento.

SOBRE O ARQUIVO

O conjunto de dados é separados em duas pastas, teste e treinamento, nelas encontramos os dados de X, y e subject. Possui ainda uma subpasta (Inertial Signals) com os dados brutos.

METODOLOGIA

Primeiramente, a análise exploratória dos dados foi realizada, examinando o comportamento das variáveis e a existência de correlações entre elas. Para melhorar a eficiência, as variáveis mais relevantes foram selecionadas e, subsequente a isso, uma redução de dimensionalidade foi efetuada utilizando a técnica de PCA. Após a preparação dos dados, o algoritmo K-MEANS foi implementado para realizar o agrupamento dos dados em clusters. O número ideal de clusters foi determinado por meio dos métodos do cotovelo (*Elbow Method*) e do *Silhouette Score*. Adicionalmente, a normalização dos dados foi aplicada para garantir que todas as variáveis contribuíssem igualmente para o modelo.

Utilizamos o método do cotovelo para identificar o ponto em que a inércia diminui drasticamente, enquanto o *Silhouette Score* foi empregado para indicar a coesão e a separação dos clusters. O número de clusters foi ajustado com base nessas métricas.

O modelo foi avaliado utilizando tanto o *Silhouette Score* quanto a inércia, além de visualizações em 2D e 3D, que foram elaboradas para interpretar os resultados de forma mais intuitiva.

RESULTADOS

MÉTRICAS DE AVALIAÇÃO:

Inércia: A inércia final do modelo foi de [valor], o que indica a compactação dos clusters.

Silhouette Score: *O Silhouette Score* médio foi de [valor], o que sugere que os clusters são bem definidos e com boa separação.

DISCUSSÃO

Os resultados mostraram que o K- MEANS foi eficaz para identificar diferentes grupos de atividades humanas, com boa separação e coesão interna dos clusters. A escolha do número de clusters foi fundamental para garantir que os agrupamentos refletissem as atividades reais registradas. No entanto, a escolha do número de componentes principais no PCA pode ter influenciado os resultados, já que a redução de dimensionalidade pode ocasionar a perda de informações importantes.

Limitações do modelo incluem a sensibilidade à escolha do número de clusters e a possibilidade de que o K- MEANS não tenha sido o melhor algoritmo para capturar a complexidade dos dados. Outras abordagens, como o K-MEANS com múltiplas execuções ou outros algoritmos de agrupamento (como DBSCAN), podem ser exploradas em projetos futuros.

CONCLUSÃO

Este projeto demonstrou a aplicação do K- MEANS no agrupamento de atividades humanas, mostrando a importância da preparação adequada dos dados e da escolha do número correto de clusters. O uso de PCA foi fundamental para reduzir a dimensionalidade e melhorar a eficiência do algoritmo

REFERÊNCIAS

Human Activity Recognition Using Smartphones. UC Irvine Machine Learning Repository. Disponível em: < https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones >

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra e Jorge L. Reyes-Ortiz. Um conjunto de dados de domínio público para reconhecimento de atividade humana usando smartphones. 21º Simpósio Europeu sobre Redes Neurais Artificiais, Inteligência Computacional e Aprendizado de Máquina, ESANN 2013. Bruges, Bélgica, 24-26 de abril de 2013.