Análise e Avaliação do Algoritmo K-means no Reconhecimento de Atividades Humanas com o Dataset HAR

Nome dos Residentes: Iaquine Santos da Silva e Renato Gomez de Sousa

Data de Entrega: 02 de dezembro de 2024

Resumo

O objetivo deste projeto foi implementar o algoritmo K-means para realizar a segmentação de atividades humanas utilizando o conjunto de dados de Reconhecimento de Atividades Humanas (HAR) com sensores de smartphones. A metodologia incluiu a extração e carregamento dos dados, análise exploratória, redução de dimensionalidade utilizando PCA, e a escolha do número ideal de clusters por meio das métricas de inércia e silhouette score. O modelo K-means foi aplicado aos dados reduzidos, utilizando o valor de K ideal encontrado. A análise revelou que o número ideal de clusters foi 2, com um silhouette score de aproximadamente 0,70, indicando uma boa separação entre os clusters. A visualização dos resultados confirmou que o algoritmo foi capaz de agrupar os dados de maneira significativa, contribuindo para a compreensão das diferentes atividades representadas no conjunto de dados.

Introdução

O reconhecimento de atividades humanas (HAR, do inglês *Human Activity Recognition*) é um campo de pesquisa interdisciplinar que busca identificar e classificar padrões de atividades realizadas por indivíduos com base em dados coletados de sensores. Com aplicações em saúde, fitness, segurança e automação residencial, o HAR tem ganhado relevância nos últimos anos devido à popularização de dispositivos vestíveis, como relógios inteligentes e smartphones equipados com acelerômetros e giroscópios. Esses dispositivos geram grandes volumes de dados multivariados, frequentemente de alta dimensionalidade, que exigem técnicas eficazes para extração de padrões úteis.

Um dos desafios no reconhecimento de atividades humanas reside na complexidade e variação intrínseca das atividades, como caminhar, correr, subir escadas

ou descansar, que podem ser influenciadas por fatores como estilo individual, ambiente ou condições de coleta dos dados. Nesse contexto, a redução de dimensionalidade e a identificação de agrupamentos são fundamentais para simplificar a análise e destacar características relevantes.

Neste projeto, optou-se pelo uso do algoritmo K-means para realizar o agrupamento dos dados. O K-means é amplamente utilizado devido à sua eficiência e simplicidade computacional, além de ser adequado para explorar padrões latentes em dados não rotulados. A aplicação do K-means possibilita:

- Redução da Complexidade: Ao agrupar os dados em clusters, conseguimos identificar semelhanças entre diferentes amostras, o que auxilia na categorização de atividades.
- Análise Exploratória Avançada: Os agrupamentos podem revelar estruturas nos dados que não seriam evidentes em uma análise convencional.
- Base para Modelagem Supervisionada: Os resultados do agrupamento podem servir como ponto de partida para modelos supervisionados, principalmente quando rótulos são escassos ou não confiáveis.

Adicionalmente, técnicas como a Análise de Componentes Principais (PCA) são empregadas para reduzir a dimensionalidade dos dados, preservando sua variância principal e permitindo uma visualização mais clara dos clusters. Essa abordagem não apenas facilita a identificação de agrupamentos significativos, mas também melhora a eficiência computacional, tornando o modelo mais adequado para lidar com grandes conjuntos de dados.

Portanto, este trabalho explora o potencial do K-means no reconhecimento de padrões de atividades humanas, utilizando o conjunto de dados *Human Activity Recognition Using Smartphones* como base para experimentação e análise.

Metodologia

- Análise Exploratória de Dados
    - Carregamento e Estruturação dos Dados
        - Extração do Dataset: O arquivo UCI HAR Dataset foi extraído utilizando a biblioteca zipfile e organizado em subpastas para dados de treino e teste.

        - Leitura dos Arquivos:
            - Dados de entrada (features) foram carregados a partir de arquivos como X_train.txt e X_test.txt.
            - Rótulos das atividades foram lidos de arquivos como y_train.txt e y_test.txt.
            - Identificadores dos sujeitos foram carregados para associar os dados a indivíduos específicos.
- Combinação dos Dados: Dados de entrada, rótulos e identificadores foram combinados em um único DataFrame para treino e teste.
    - Inspeção dos Dados
- Resumo Estatístico:
    - Comando .describe() usado para entender a distribuição das variáveis.
    - Visualização inicial das correlações entre as variáveis usando uma matriz de correlação e um heatmap.
- Distribuição das Variáveis:
    - A densidade da primeira variável foi visualizada com histograma e estimativa de densidade kernel (KDE).

- Pré-processamento
    - Normalização
        - Todas as variáveis de entrada foram escaladas usando StandardScaler, garantindo que médias fossem 0 e desvio padrão 1.
        - Esta etapa é crítica para algoritmos baseados em distâncias, como o K-means.
    - Redução de Dimensionalidade
        - PCA (Principal Component Analysis):
            - Redução para 2 componentes principais, facilitando a visualização dos clusters em gráficos bidimensionais. Variáveis reduzidas foram usadas como entrada para o K-means.
- Escolha do Número de Clusters (K)
    - Método do Cotovelo:

        - Métrica de inércia foi calculada para valores de KKK entre 2 e 10.
        - O ponto de inflexão no gráfico foi usado como critério inicial.
- Silhouette Score:
        - Métrica adicional usada para validar a coesão e separação dos clusters.
        - K com maior valor de Silhouette Score foi escolhido como ideal.

- Aplicação do K-means
    - Inicialização
        - O algoritmo foi inicializado com KMeans++, para garantir a escolha eficiente de centroides iniciais.
    - Execução e Clusterização
        - Clusters foram gerados a partir dos dados transformados pelo PCA.
        - Labels de cluster foram atribuídos a cada observação.
    - Avaliação dos Clusters
        - Estabilidade: Centroides foram inspecionados para verificar convergência.
    - Visualização
        - Clusters foram visualizados em um gráfico de dispersão 2D, usando as componentes principais.
    - Métrica de Avaliação:
        - Silhouette Score calculado para o agrupamento final foi 0.7010.7010.701, indicando boa separação entre clusters.

Resultados

Métricas de Avaliação

- Número ideal de clusters (K): Após a análise utilizando o método do cotovelo e o Silhouette Score, determinou-se que o número ideal de clusters é 2, sendo este o valor que maximiza a coesão intra-cluster e a separação inter-cluster.
- Silhouette Score: O Silhouette Score médio para o modelo foi 0.701, indicando uma qualidade satisfatória dos clusters formados. Este valor reflete a proximidade dos pontos aos seus próprios clusters e a distância deles em relação aos clusters vizinhos, apontando para boa separação e coesão.

- Centroides dos Clusters:
    - Centróide 1: [−13.98,1.02][-13.98, 1.02][−13.98,1.02]
    - Centróide 2: [17.24,−1.26][17.24, -1.26][17.24,−1.26] Esses valores são as médias das componentes principais atribuídas a cada cluster, após a redução de dimensionalidade.

- Visualizações Gráficas
    - Análise de Determinação de K:
        - Método do Cotovelo: O gráfico de inércia (variação explicada) mostrou uma redução acentuada em K=2K=2K=2, confirmando a escolha de 2 clusters como ideal.
        - Silhouette Score: O gráfico indicou que K=2K=2K=2 também apresentou o maior valor de silhouette score, reforçando a escolha.
- Visualização dos Clusters Após PCA:
    - A dispersão dos clusters foi plotada no espaço bidimensional das duas principais componentes do PCA. Cada cluster foi representado com cores distintas:
        - Cluster 1: Distribuído principalmente na parte inferior esquerda.
        - Cluster 2: Concentrado na parte superior direita.
        - Essa visualização confirma que os clusters possuem boa separação, com pouca sobreposição.

Discussão

Durante o desenvolvimento do modelo K-means aplicado ao dataset Human Activity Recognition Using Smartphones (HAR), várias decisões técnicas e metodológicas influenciaram os resultados obtidos, os quais são discutidos criticamente a seguir. O número de clusters ideal, determinado pelo método do Silhouette Score, indicou que dois clusters eram adequados. Contudo, esse resultado não captura a granularidade esperada, considerando que o dataset HAR contém seis categorias de atividades. Isso sugere que a redução de dimensionalidade com PCA, que utilizou apenas duas componentes principais, pode ter simplificado excessivamente as informações. O Silhouette Score médio de 0.701 indica uma separação razoável entre os clusters, mas

sem uma distinção excepcional, refletindo que o K-means conseguiu formar grupos coesos, porém com uma sobreposição significativa entre os dados.

A análise pós-PCA forneceu uma representação clara dos clusters em um espaço bidimensional, facilitando a interpretação e identificação de padrões, mas essa representação é limitada e não captura a verdadeira complexidade dos dados originais. Em termos de limitações do modelo, a escolha de usar apenas duas componentes principais via PCA pode ter reduzido a variabilidade explicada, omitindo informações cruciais e impactando diretamente a capacidade do modelo de identificar padrões complexos presentes nas 561 dimensões originais.

Embora o Silhouette Score seja uma métrica robusta para avaliar a coesão e separação, ele pode ser insuficiente para dados com clusters não esféricos ou densidades variadas, como no caso dos dados de sensores do HAR. O algoritmo K-means, que assume clusters esféricos e de tamanhos semelhantes, pode não ser adequado para o contexto do HAR, onde os dados representam atividades humanas com padrões mais complexos. A necessidade de definir o número de clusters previamente limita a flexibilidade do K-means, e apesar da análise com diferentes valores de K, a escolha final pode não refletir a verdadeira estrutura dos dados. Além disso, a escalabilidade do algoritmo, embora mitigada pela normalização com StandardScaler, ainda pode ser um desafio em aplicações com maior volume de dados. Em relação ao impacto das escolhas feitas, o pré-processamento com StandardScaler e a remoção de colunas duplicadas foram passos cruciais para garantir resultados mais confiáveis, mas escolhas adicionais, como a seleção de variáveis relevantes para a clusterização, poderiam ter melhorado os resultados.

A combinação dos métodos do Cotovelo e Silhouette para determinar o número ideal de clusters foi adequada, mas a dependência de apenas duas métricas pode não ter sido suficiente para capturar toda a complexidade do dataset. A implementação do PCA, embora útil para a visualização, comprometeu a retenção de variabilidade, limitando o desempenho potencial do modelo. Por fim, embora as representações visuais dos clusters tenham facilitado a interpretação, elas podem ser enganosas, pois não consideram as dimensões perdidas na redução de dados. Esta análise destaca as limitações dos métodos tradicionais, como o K-means, quando aplicados a dados complexos, e sugere caminhos para uma compreensão mais profunda e representativa da estrutura presente no dataset HAR.

Conclusão e Trabalhos Futuros

O projeto de implementação do algoritmo K-means para o dataset de Reconhecimento de Atividade Humana (HAR) forneceu insights valiosos sobre o comportamento dos dados e a aplicação de técnicas de clustering. A análise exploratória mostrou a complexidade do conjunto de dados, que possui 561 características, exigindo uma abordagem de redução de dimensionalidade para facilitar a análise. O uso do PCA (Análise de Componentes Principais) permitiu uma visualização mais clara e compreensível dos dados em 2D, possibilitando a aplicação do K-means de forma mais eficiente.

A escolha do número ideal de clusters, através dos métodos do Cotovelo e Silhouette Score, indicou que o modelo com 2 clusters foi o mais adequado. A visualização dos clusters pós-PCA e a avaliação das métricas de estabilidade, como o Silhouette Score (0.70), indicam que o modelo tem um desempenho razoável para separar os dados em grupos coerentes.

Embora o modelo tenha demonstrado uma boa capacidade de clustering, existem várias áreas que podem ser aprimoradas para aumentar a precisão e a eficiência do modelo:

- Aprimoramento da escolha de K: O método de escolha de K pode ser explorado mais a fundo, utilizando técnicas como o método da média Silhouette, ou aplicando validação cruzada para garantir que a escolha do número de clusters não seja excessivamente otimista.
- Uso de outras técnicas de redução de dimensionalidade: Embora o PCA tenha sido eficaz, outras técnicas de redução de dimensionalidade, como t-SNE ou UMAP, podem ser exploradas para verificar se oferecem uma melhor separação dos clusters e facilitam a visualização dos dados.
- Exploração de diferentes algoritmos de clustering: Além do K-means, algoritmos como DBSCAN ou HDBSCAN podem ser testados para comparar com a segmentação realizada pelo K-means, especialmente em dados com formas de distribuição não esféricas ou com densidades variadas.
- Avaliação do impacto de pré-processamento: O impacto de diferentes técnicas de normalização e transformação dos dados (como MinMax Scaling ou robust scaling) pode ser investigado para verificar se há uma melhoria na formação dos clusters.

- Análise mais profunda dos clusters: Após a segmentação, uma análise mais detalhada dos clusters gerados poderia ajudar a entender melhor os grupos de atividades e as características que os diferenciam, podendo abrir caminho para um modelo de classificação supervisionada mais robusto.

Esses aprimoramentos podem fornecer uma visão mais precisa das atividades humanas, melhorando a eficácia do modelo no reconhecimento de padrões complexos.

Referências

- ZHENG, Alice; CASARI, Amanda. **Feature engineering for machine learning**: principles and techniques for data scientists. 1. ed. Sebastopol, CA: O'Reilly Media, 2018. ISBN 978-1-491-95324-2.
- HAPKE, Hannes; NELSON, Catherine. **Building machine learning pipelines**: automating model life cycles with TensorFlow. 1. ed. Beijing; Boston; Farnham; Sebastopol; Tokyo: O'Reilly Media, 2020. ISBN 978-1-492-05319-4.
- AVILA, Julian; HAUCK, Trent. **Scikit-learn cookbook**: over 80 recipes for machine learning in Python with scikit-learn. 2. ed. Birmingham; Mumbai: Packt Publishing, 2017. ISBN 978-1-78728-638-2.
- AVILA, Julian; HAUCK, Trent. **Scikit-learn cookbook**: over 80 recipes for machine learning in Python with scikit-learn. 2. ed. Birmingham; Mumbai: Packt Publishing, 2017. ISBN 978-1-78728-638-2.
- DANGETI, Pratap. **Statistics for machine learning**: build supervised, unsupervised, and reinforcement learning models using both Python and R. Birmingham; Mumbai: Packt Publishing, 2017. ISBN 978-1-78829-575-8.
- PATEL, Ankur A. **Hands-on unsupervised learning using Python**: how to build applied machine learning solutions from unlabeled data. 1. ed. Beijing; Boston; Farnham; Sebastopol; Tokyo: O'Reilly Media, 2019. ISBN 978-1-492-03564-0.
- JHA, Suraj. **Top Instagram influencers data cleaned**. Disponível em: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned. Acesso em: 17 de novembro de 2024.