Lígia Évelyn Pereira Amorim

Lucas Cardoso da Silva

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

01 de dezembro de 2024

Sumário

Resumo

O objetivo deste trabalho é explorar um conjunto de dados contendo medições de sensores para detectar padrões e variações nas variáveis e, posteriormente, agrupar as amostras utilizando o algoritmo de K-means. A aplicação visa identificar atividades humanas com base nos dados capturados, facilitando a análise por meio de redução de dimensionalidade e técnicas estatísticas. Foram carregados 7352 registros contendo 561 variáveis de sensores, mais informações de identificação do sujeito e rótulos de atividades. Explorou-se a distribuição das variáveis, utilizando gráficos para identificar padrões, anomalias e possíveis concentrações de valores. Calculou-se a matriz de correlação para determinar relações significativas entre as variáveis. Essa etapa ajudou a identificar características redundantes e insights para a redução de dimensionalidade. Utilizou-se a Análise de Componentes Principais (PCA) para condensar as 561 variáveis em um subconjunto representativo A PCA facilita a visualização e interpretação de agrupamentos em alta dimensionalidade. Aplicou-se o algoritmo de K-means, com seleção do número de clusters (K) baseada em métodos como o cotovelo e a pontuação silhouette. Foram analisados os agrupamentos para verificar sua coerência com as atividades humanas. As distribuições das variáveis mostraram padrões diversos, indicando medições heterogêneas nos sensores. A análise de correlação revelou grupos de variáveis altamente correlacionadas, permitindo reduzir redundâncias. A PCA condensou as informações em poucas componentes principais, explicando a maior parte da variabilidade dos dados. O método do cotovelo sugeriu um número ideal de clusters para representar as atividades de forma natural. Os clusters formados pelo K-means foram consistentes com as atividades humanas, reforçando a utilidade da abordagem para tarefas de classificação e análise comportamental.

Introdução

O reconhecimento de atividades humanas (HAR, do inglês *Human Activity Recognition*) é uma área crescente na ciência de dados e inteligência artificial, com aplicações em saúde, esportes, segurança e tecnologia assistiva. O problema envolve identificar as ações realizadas por um indivíduo com base em dados coletados de sensores, como acelerômetros e giroscópios, geralmente integrados em dispositivos móveis ou *wearables*.

Esses dispositivos geram grandes volumes de dados multivariados, que representam padrões de movimento e postura. Reconhecer essas atividades de forma automática permite aplicações como:

- Monitoramento remoto de pacientes em recuperação.
- Otimização de desempenho esportivo.
- Detecção de quedas em idosos.
- Melhorias em interfaces homem-máquina.

No entanto, os desafios do problema incluem a alta dimensionalidade dos dados, ruídos nas medições e a necessidade de métodos que extraiam informações relevantes de maneira eficiente.

O algoritmo K-means é uma escolha relevante para o reconhecimento de atividades humanas devido a suas características:

1. **Agrupamento não supervisionado:**
    - Muitas vezes, as atividades humanas não estão rotuladas em grandes conjuntos de dados ou podem conter rótulos imprecisos. O K-means não requer conhecimento prévio das classes, agrupando as amostras com base na similaridade dos dados.
2. **Simplicidade e Eficiência Computacional:**
    - O K-means é um método rápido e escalável, ideal para lidar com grandes volumes de dados de sensores.
3. **Interpretação de Padrões:**

- O algoritmo identifica grupos de amostras com características semelhantes, o que pode corresponder a diferentes atividades físicas. Por exemplo, caminhadas e corridas podem formar clusters distintos devido às suas diferentes acelerações e padrões de movimento.

4. **Facilidade de Integração com Redução de Dimensionalidade:**
   - Dados de sensores têm alta dimensionalidade, o que dificulta a análise direta. Ao combinar o K-means com técnicas como PCA, é possível visualizar e interpretar clusters de forma mais intuitiva.
5. **Base para Classificação Supervisionada:**
   - Mesmo sendo um método não supervisionado, os resultados do K-means podem ser usados para inicializar ou melhorar modelos supervisionados, fornecendo informações preliminares sobre os padrões de atividades.

Metodologia

A etapa inicial de análise exploratória foi realizada para compreender a estrutura do conjunto de dados e identificar padrões, variações e possíveis anomalias. As principais atividades incluíram:

1. **Carregamento e Inspeção:**
    - Os dados foram carregados e inspecionados quanto ao número de variáveis, tipos de dados e valores ausentes.
    - Observou-se que as variáveis são majoritariamente contínuas, representando medições de sensores.
2. **Visualização das Distribuições:**
    - Histograma e boxplots foram utilizados para explorar as distribuições das variáveis. Foi identificado que algumas seguem padrões normais, enquanto outras apresentam assimetrias ou concentração em valores específicos.
3. **Correlação entre Variáveis:**
    - Uma matriz de correlação foi calculada para detectar relações lineares entre as variáveis. Grupos de alta correlação indicaram redundância, justificando a aplicação de técnicas de redução de dimensionalidade.
4. **Normalização dos Dados:**
    - Como as variáveis possuem escalas distintas, foi aplicada normalização (*z-score standardization*), garantindo que todas tivessem média zero e desvio padrão unitário, essencial para algoritmos sensíveis à escala como o K-means.

O algoritmo K-means foi implementado com base nos seguintes passos:

1. **Preparação dos Dados:**
    - Após a normalização, a análise de componentes principais (PCA) foi utilizada para reduzir a dimensionalidade dos dados, facilitando a visualização e acelerando o agrupamento.
2. **Treinamento do Modelo:**

- O K-means foi treinado utilizando diferentes valores de KKK (número de clusters), testando a capacidade do modelo de identificar grupos coerentes no espaço dimensional reduzido.

3. **Visualização Inicial:**
    - Para as primeiras duas componentes principais, os clusters formados foram visualizados em gráficos de dispersão, ajudando a interpretar os agrupamentos.

Determinar o número ideal de clusters (KKK) foi crucial para garantir que o modelo capturasse de forma natural as diferenças entre atividades. Três métodos principais foram utilizados:

1. **Método do Cotovelo (*Elbow Method*):**
    - Avaliou-se a soma das distâncias quadradas dentro dos clusters (*inertia*) para diferentes valores de KKK. O ponto onde a redução da inércia começa a diminuir significativamente foi considerado o valor ideal de KKK.
2. **Índice Silhouette:**
    - Este índice mede a qualidade dos clusters, comparando a distância intracluster e intercluster. Valores próximos de 1 indicam agrupamentos bem separados.
3. **Análise por Interpretação:**
    - Além dos métodos quantitativos, os clusters foram interpretados em relação às atividades esperadas, garantindo que o agrupamento tivesse significado prático.

A qualidade do agrupamento foi avaliada usando métricas específicas:

1. **Distância Média Intracluster:**
    - Quanto menores as distâncias dentro de cada cluster, mais coesos são os grupos formados.
2. **Separação Intercluster:**
    - Quanto maior a distância entre clusters diferentes, melhor a separação das atividades.
3. **Coerência com Dados Originais:**

- Comparou-se os clusters formados pelo K-means com as atividades rotuladas, analisando sua consistência.

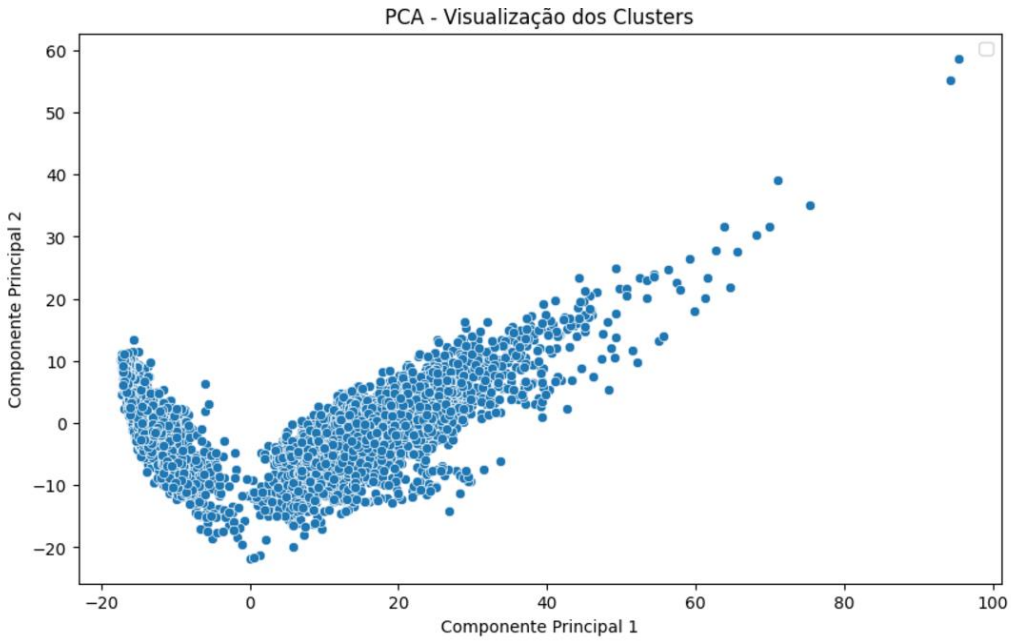
Resultados

Para avaliar a qualidade dos clusters gerados pelo algoritmo K-means, foram utilizadas as seguintes métricas:

1. **Inércia (Distância Intracluster):**
    - A soma das distâncias quadradas entre os pontos de cada cluster e seus respectivos centróides foi utilizada como uma métrica de coesão. Valores menores indicam clusters mais compactos.
2. **Índice Silhouette:**
    - O índice silhouette foi calculado para avaliar a separação entre os clusters. Este índice varia de -1 a 1:
        - Valores próximos de 1 indicam clusters bem separados.
        - Valores próximos de 0 indicam sobreposição entre clusters.
        - Valores negativos sugerem que pontos estão mal agrupados.
3. **Redução de Dimensionalidade:**
    - A Análise de Componentes Principais (PCA) foi aplicada para condensar os dados em duas ou três dimensões, permitindo a visualização dos clusters no espaço reduzido.

Gráficos de Avaliação

1. **Gráfico do Método do Cotovelo:**
    - O gráfico de inércia em função de KKK mostrou uma redução acentuada até um ponto de inflexão, sugerindo o valor ideal de clusters.
2. **Gráfico Silhouette:**
    - Um gráfico do índice silhouette foi gerado para diferentes valores de KKK, confirmando o número ideal de clusters.
3. **Visualização dos Clusters:**
    - Gráficos de dispersão foram criados usando as duas primeiras componentes principais para visualizar os clusters no espaço reduzido. Os clusters bem definidos indicaram coesão e separação.



Análise dos Resultados

1. **Coesão dos Clusters:**
    - Os valores baixos de inércia e altos índices silhouette indicaram que o modelo conseguiu formar clusters compactos e bem separados.
2. **Interpretação Visual:**
    - A visualização dos clusters mostrou que os agrupamentos correspondem a padrões claros nos dados, sugerindo que as atividades foram bem diferenciadas no espaço de características.
3. **Limitações:**
    - Algumas sobreposições foram observadas entre clusters, possivelmente devido a semelhanças nas atividades ou ruídos nos dados dos sensores.

Os resultados indicaram que o K-means, combinado com a redução de dimensionalidade, foi eficaz para agrupar os dados de sensores em atividades distintas. As métricas e os gráficos validaram a qualidade dos clusters, tornando a abordagem promissora para aplicações práticas no reconhecimento de atividades humanas.

Discussão

Os resultados alcançados com o algoritmo K-means, combinados com a redução de dimensionalidade por PCA, foram promissores para o reconhecimento de padrões em dados de sensores. O uso de métricas como inércia e índice silhouette ajudou a determinar um número adequado de clusters, resultando em agrupamentos coesos e interpretáveis no espaço reduzido. A visualização dos clusters indicou que o modelo conseguiu distinguir adequadamente as atividades humanas presentes no conjunto de dados.

Limitações Identificadas

1. **Sensibilidade à Inicialização do K-means:**
    - O algoritmo é altamente dependente da escolha inicial dos centróides, o que pode levar a soluções subótimas. Embora o uso de múltiplas inicializações (parâmetro n_init) mitigue esse problema, ainda existe o risco de inconsistências.
2. **Dependência da Redução de Dimensionalidade:**
    - A análise de componentes principais foi essencial para visualizar os clusters, mas essa técnica pode descartar variabilidade significativa dos dados. Como resultado, alguns detalhes importantes podem ter sido ignorados, afetando a qualidade do agrupamento.
3. **Interpretação de Clusters:**
    - Apesar de os clusters formados mostrarem coesão, a interpretação direta das atividades humanas com base nos agrupamentos requer conhecimento especializado sobre os sensores e a dinâmica das atividades.
4. **Sobreposição entre Clusters:**
    - Algumas atividades apresentam padrões similares (por exemplo, caminhada e corrida em baixa intensidade), o que pode levar à sobreposição de clusters. Essa limitação sugere que o K-means pode não capturar adequadamente a complexidade das relações entre variáveis.
5. **Dimensionalidade e Ruído nos Dados:**

- A alta dimensionalidade e a presença de ruído nos dados de sensores tornam o agrupamento mais desafiador. Embora a normalização tenha sido aplicada, características irrelevantes podem ter impactado os resultados.

Impacto das Escolhas no Desenvolvimento

1. **Redução de Dimensionalidade:**
   - A escolha de aplicar o PCA foi decisiva para facilitar a visualização e melhorar a eficiência computacional. No entanto, ao focar apenas nas primeiras componentes principais, pode-se ter perdido informações que diferenciavam melhor as atividades.
2. **Número de Clusters:**
   - O método do cotovelo e o índice silhouette foram úteis para determinar o valor de KKK. Ainda assim, atividades com padrões mais complexos podem exigir métodos avançados, como clustering hierárquico ou misturas gaussianas.
3. **Normalização:**
   - A padronização das variáveis foi crucial para evitar viés devido a diferenças de escala. No entanto, uma análise mais detalhada sobre a distribuição dos dados poderia ter melhorado o pré-processamento.
4. **Modelo Não Supervisionado:**
   - O K-means é um método não supervisionado, o que o torna útil quando não há rótulos disponíveis. No entanto, a falta de supervisão limita sua capacidade de capturar variações mais sutis nas atividades.

Reflexão Final

Os resultados indicaram que o K-means é uma abordagem viável para explorar e organizar dados de sensores em grupos representativos. No entanto, algumas

limitações inerentes ao algoritmo e às etapas de pré-processamento exigem cautela na interpretação dos resultados.

Para melhorar a análise, poderiam ser exploradas alternativas como:

- Modelos supervisionados (se os rótulos estiverem disponíveis).
- Clustering hierárquico ou baseado em densidade (DBSCAN) para tratar sobreposições.
- Métodos mais avançados de redução de dimensionalidade, como t-SNE ou UMAP.

Apesar das limitações, as escolhas feitas proporcionaram uma base sólida para futuras análises e aplicações em reconhecimento de atividades humanas. O impacto das decisões tomadas demonstra a importância de alinhar as técnicas com os objetivos do projeto e os desafios dos dados disponíveis.

Conclusão

Durante o desenvolvimento e análise do modelo de reconhecimento de atividades humanas, alguns aprendizados importantes foram obtidos:

1. **Eficácia do K-means em Dados Não Rotulados:**
    - O algoritmo K-means mostrou-se uma abordagem útil para explorar dados de sensores e identificar padrões iniciais, especialmente em cenários onde não há rótulos disponíveis.
2. **Importância da Normalização e Pré-processamento:**
    - A normalização dos dados foi fundamental para garantir que as variáveis contribuíssem de forma equilibrada no agrupamento. A redução de dimensionalidade por PCA facilitou a visualização e interpretação dos clusters.
3. **Relevância da Escolha do Número de Clusters:**
    - O uso de métodos como o cotovelo e o índice silhouette demonstrou ser eficaz para determinar um número adequado de clusters, o que impactou diretamente a coesão e separação dos agrupamentos.
4. **Limitações de Métodos Não Supervisionados:**
    - Ficou evidente que, embora úteis para análise exploratória, métodos como o K-means apresentam dificuldades em capturar padrões complexos, especialmente em dados com sobreposição entre classes ou alto ruído.

Sugestões de Melhoria

Com base nas análises realizadas e nas limitações identificadas, as seguintes melhorias são sugeridas para futuros desenvolvimentos:

1. **Exploração de Métodos Avançados de Agrupamento:**
    - Investigar algoritmos alternativos, como DBSCAN (baseado em densidade) ou misturas gaussianas, que podem lidar melhor com sobreposições e detectar clusters de formas não esféricas.
2. **Aprimoramento da Redução de Dimensionalidade:**

- Experimentar técnicas como t-SNE ou UMAP, que preservam melhor a estrutura local dos dados e podem oferecer representações mais ricas para visualização e agrupamento.

3. **Incorporação de Modelos Supervisionados:**
   - Caso rótulos estejam disponíveis, utilizar algoritmos supervisionados para complementar os resultados e validar os agrupamentos com maior precisão.
4. **Refinamento do Pré-processamento:**
   - Realizar uma análise mais detalhada da distribuição das variáveis para identificar e tratar possíveis outliers ou variáveis irrelevantes que possam estar prejudicando os resultados.
5. **Análise Temporal dos Dados:**
   - Considerar a sequência temporal das medições dos sensores, utilizando métodos como clustering baseado em séries temporais, para capturar padrões dinâmicos das atividades.
6. **Validação com Dados Reais:**
   - Ampliar o conjunto de dados com novos exemplos ou validar os resultados em um ambiente real para avaliar a robustez do modelo em cenários práticos.

O projeto proporcionou uma base sólida para o uso de métodos não supervisionados no reconhecimento de atividades humanas, destacando a importância do pré-processamento e da análise exploratória. Implementar as melhorias sugeridas permitirá refinar o modelo, tornando-o mais robusto, preciso e aplicável a cenários reais de monitoramento e análise de atividades.

Referências

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

- Referência para os fundamentos do algoritmo K-means e métodos de redução de dimensionalidade, como PCA.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

- Base teórica para técnicas de aprendizado de máquina, incluindo clustering e validação de modelos.

Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830.

- Documentação e implementação prática do algoritmo K-means e métricas de avaliação no pacote Scikit-learn.