**CEPEDI**

**CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS**

Holiver Nicolas Moura Casé

Tarcisio Lopes de Almeida Sousa

**Relatório Técnico: Reconhecimento de Atividades Humanas Utilizando K-Means e PCA**

Juazeiro-BA

2024

**Holiver Nicolas Moura Casé**

**Tarcisio Lopes de Almeida Sousa**

**Relatório Técnico: Reconhecimento de Atividades Humanas Utilizando K-Means e PCA**

Relatório apresentado como parte dos requisitos para obtenção da nota final da 10° unidade da disciplina de trilha de Ciência de Dados

Juazeiro-BA

2024

**RESUMO**

Este trabalho investiga o agrupamento de atividades humanas utilizando técnicas de aprendizado de máquina não supervisionado, com foco no algoritmo K-Means e redução de dimensionalidade por PCA. O estudo baseia-se no UCI HAR Dataset, onde foram identificados padrões e avaliados clusters em diferentes configurações. A análise destacou a aplicabilidade do método do cotovelo, silhouette score, Davies-Bouldin score e Calinski-Harabasz Score para determinar o número ideal de clusters, revelando insights sobre a separação entre as atividades.

**SUMÁRIO**

**1. INTRODUÇÃO......................................................................................5**

**2. METODOLOGIA...................................................................................6**

**3 DISCUSSÃO..........................................................................................7**

**4. CONCLUSÃO E TRABALHOS FUTUROS..........................................8**

**5. REFERÊNCIAS....................................................................................9**

**INTRODUÇÃO**

A classificação de atividades humanas é uma área crescente na interseção entre tecnologia e saúde, com aplicações que vão desde monitoramento de pacientes até dispositivos vestíveis inteligentes. O desafio reside na alta dimensionalidade dos dados captados por sensores, que dificulta a identificação de padrões. Este estudo aborda a questão utilizando redução de dimensionalidade (PCA) para projeção visual e agrupamento não supervisionado (K-Means), oferecendo uma solução prática para explorar a estrutura latente dos dados.

**1 METODOLOGIA**

1. Coleta e Processamento dos Dados:
   * O conjunto de dados UCI HAR Dataset foi carregado, incluindo medições captadas por acelerômetros e giroscópios de dispositivos móveis.
   * Os dados foram limpos e padronizados com StandardScaler para garantir uniformidade entre as variáveis.
2. Redução de Dimensionalidade:
   * Foi aplicada a técnica de Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados originais e facilitar a visualização.
   * Os dois primeiros componentes principais explicaram a maior parte da variância, permitindo uma análise visual dos agrupamentos.
3. Clustering com K-Means:
   * O algoritmo K-Means foi utilizado para formar clusters, variando o número K de 2 a 10.
   * Foram aplicadas métricas para escolha do número ideal de clusters:
     1. Método do Cotovelo: Para avaliar a redução da inércia dentro dos clusters.
     2. Silhouette Score: Para medir a separação e coesão entre os clusters formados.
     3. Davies-Bouldin score: Para medir a dispersão dos clusters.
     4. Calinski-Harabasz Score: Para avaliar a densidade e a separação dos clusters.
4. Validação dos Resultados:
   * Realizou-se múltiplas inicializações do algoritmo com o método k-means++, garantindo a consistência dos agrupamentos.
   * Resultados foram visualizados em gráficos 2D e 3D após a aplicação do PCA.
   * Distribuição de atividades de aspectos semelhantes dentro de cada cluster.

**2 DISCUSSÃO**

1. Equilíbrio entre Classes:
   * Embora o modelo não utilize rótulos, o silhouette score indicou bom desempenho geral para K=4K = 4K=4 e K=5K = 5K=5, sugerindo que essas configurações respeitam a separação natural entre as atividades.
2. Desafios com Certos Clusters:
   * Observou-se que atividades fisicamente similares (e.g., andar e subir escadas) apresentaram maior sobreposição, indicando dificuldade do algoritmo em diferenciá-las.
3. Fatores que Influenciam o Agrupamento:
   * A qualidade das medições e o desbalanceamento entre classes provavelmente impactaram a consistência dos clusters.
   * Ruídos nos dados podem ter causado menor coesão nos agrupamentos.
4. Possíveis Melhorias:
   * Ajustar hiperparâmetros como o número de inicializações.
   * Implementar técnicas avançadas de seleção de features ou incorporar algoritmos alternativos, como DBSCAN, para lidar com dados ruidosos.
5. Reflexões Adicionais:
   * A aplicação do PCA foi crucial para destacar padrões que, de outra forma, permaneceriam escondidos nos dados originais.
   * A pouca variação do silhouette score em diferentes K’s sugere baixa separação em certas regiões do espaço de características.

**3** **CONCLUSÃO E TRABALHOS FUTUROS**

O estudo evidenciou a eficácia do K-Means em conjunto com o PCA na análise de agrupamentos para reconhecimento de atividades humanas. A metodologia forneceu uma visão clara das relações entre as atividades, ainda que limitada por ruídos e desbalanceamento do dataset. Melhorias futuras incluem a coleta de dados mais equilibrados, o uso de técnicas supervisionadas para comparação e ajustes na arquitetura do modelo para otimizar a separação entre clusters.

**4. 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.

2 Scikit-learn Documentation. Disponível em: <https://scikit-learn.org>.

3 UCI Machine Learning Repository. Human Activity Recognition Dataset. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>.