Solving Clustering Problem

Report

Course: **CPS 803**
Machine Learning
Semester: **F2024**
Instructor: Elodie Lugez

Prepared by-
Parnia Zare

TABLE OF CONTENTS Page number

Background

I chose the The MHEALTH (Mobile HEALTH) dataset which I came across from the UCI Machine Learning Repository. I found this dataset interesting after reading about how the data was recorded from body motions and vital signs recordings of ten volunteers engaged in twelve distinct physical activities. This data was captured using Shimmer wearable sensors. These sensors were placed on the chest, right wrist, and left ankle of each volunteer and recorded data such as acceleration, rate of turn, and magnetic field orientation, along with 2-lead ECG measurements. The dataset provides a detailed view of various physical motions, from static postures to dynamic activities. It is appropriate for clustering analysis to find patterns in physical activities based on sensor data, which can be applied to real-world scenarios including activity recognition and health monitoring.

Recorded at a sampling rate of 50 Hz, the sensors provide a significant insight on human motion, highlighting subtle differences across a spectrum of activities performed under realistic, non-laboratory conditions. This approach not only enhances the data's applicability to daily health and fitness monitoring but also helps in generalizing findings to common activities of daily living.

This data collection facilitates the exploration of activity clustering, aiming to identify and analyze patterns that emerge from the physiological responses and movement dynamics measured by the sensors. Each activity was precisely recorded and labeled, providing a solid foundation for detailed analytical studies aimed at improving our understanding of human physical behavior in diverse conditions.

Methods

1. **Data Loading**:
   - Data is loaded into a Python environment using pandas. This step includes specifying column names manually which were given.
   - A subset of 1000 rows is first loaded to streamline the development and testing of the processing pipeline.
2. **Data Cleaning**:
   - **Missing Values**: Rows containing any missing data are removed to ensure the quality and consistency of the dataset.
   - **Outlier Detection**: Outliers are identified using the statistical Z-score method. Any data point with a Z-score greater than 3 in any dimension is chosen as an

outlier and removed from analysis. This helps reduce the impact of extreme values on the clustering process.

3. **Standard Scaling**:
   - Using StandardScaler from scikit-learn I standardized the features by removing the mean and scaling to unit variance. This normalization is improtant as it makes sure that all features are equally influencing the analysis, avoiding any single feature with larger scale from controlling the distance calculations used in clustering.
4. **PCA (Principal Component Analysis)**:
   - Dimensionality is reduced to two principal components using PCA. This helps in reducing the computational load and also enhances visualizing the dataset.

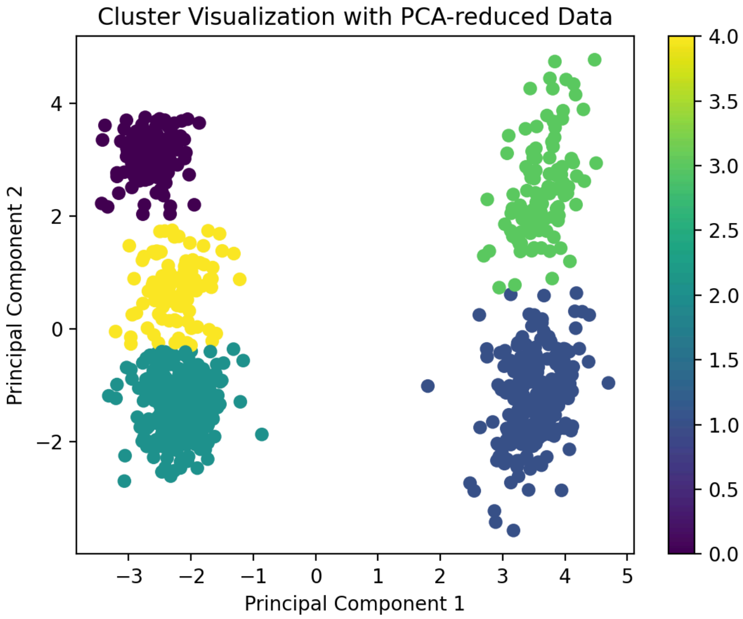
Clustering

5. **KMeans Clustering**:
   - The KMeans algorithm from scikit-learn is utilized to the 2D data to partition it into five clusters. This number was picked based on preliminary analysis and the need to balance granularity with overfitting.
   - The clustering process iterates up to 300 times or until merging, whichever occurs first.
6. **Silhouette Score**:
   - After clustering, the silhouette score is calculated to test the effectiveness of the clustering. This metric demonstrates how similar an object is to its own cluster compared to other clusters.
7. **Visualization**:
   - The clusters are visualized using a scatter plot with the two principal components on the axes and the cluster assignments represented by different colors. This visualization helps us observe the spatial distribution of clusters and their separation.

Results

A total of 100 outliers were detected and removed from the dataset. The preprocessing steps' purpose was to refine the dataset, ensuring that the following clustering was not negatively affected by large values that could distort the analysis.

The clustering achieved a silhouette score of 0.66, showing a reasonably clear separation between the clusters. This score suggests that the clusters are well-defined and distinct from each other, with data points within each cluster being more similar to each other than to points in other clusters. We can see this using a 2D scatter plot derived from the principal component analysis (PCA). This visualization helps illustrate how the dataset divides into distinct groups based on the physical activities recorded by the sensors.



The visualization and the high silhouette score together suggest that the chosen features and preprocessing methods effectively captured the underlying patterns in the dataset. Evidently, each cluster represents distinct physical activity categories, each of which has distinct movement patterns and intensities recorded by the sensors.

Conclusion

Different groupings of physical activities were identified through the use of clustering techniques in the examination of the MHEALTH dataset. With a noteworthy silhouette score of 0.66, the clusters were clearly defined, indicating that preprocessing, dimensionality reduction, and grouping were successful in capturing the activity patterns.

References

Analysis of Physical Activities Using Wearable Sensor Data via Machine Learning Approaches." *MHEALTH Dataset*, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/MHEALTH