DATA MINING REPORT

Clustering Based Anomaly Detection & K-means Clustering

Akshay Chauhan | 23055738

Link to Human Activity Recognition using smartphoes (UCI) Dataset

INTRODUCTION

This report presents the comparison of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and K-Means-based distance outlier detection methods uses Principal Component Analysis (PCA) for dimensionality reduction on the Human Activity Recognition (HAR) dataset from UCI. The visualization displays the results of two anomaly detection methods and cluster structure identification to illustrate their positive and negative aspects for performing discovery in the same two-dimensional space. The goal is to apply (DBSCAN) and K-means Clustering to identify natural groupings and detect anomalies in the data. The main responsibility of data mining with machine learning includes anomaly detection which finds deviations that significantly deviate from typical patterns.

Dataset Description: The dataset consists of sensor measurements (accelerometer and gyroscope) collected from smartphones with 30 subjects performing various activity such as **walking, sitting, standing, walking upstairs, walking downstairs, and laying** with 7352 samples and 561 features.

DATA PRE-PROCESSING TASKS APPLIED ON THE DATASET

- The dataset is standardized using StandardScaler to ensure all features has 0 mean, standard deviation 1 and unit variance.
- **Principal Component Analysis (PCA)** was applied to reduce dimensionality, retaining the first two principal components for visualization.

COMPARISITON AND ANAYLSIS

DBSCAN:

- Using DBSCAN enables anomaly detection through low-density area detection within the dataset. The plot highlights anomalies by using the red color for points marked as **-1** (red). The transformation based on PCA aids in identifying unusual observations when they exist independently or at low-density intervals in the restructured space.
- With DBSCAN, the number of clusters does not need specification because the parameters **eps** and min_samples identify cluster boundaries.

K-MEANS CLUSTERING:

- The K-Means algorithm served this study to locate outliers through cluster centroid distance detection. The method detects outliers through distances exceeding defined threshold values using peach-colored label **1** for such points whereas blue label **0** indicates normal data. The K-Means clustering technique requires data points to form round clusters with uniform spread which does not always match irregular real-world data patterns.

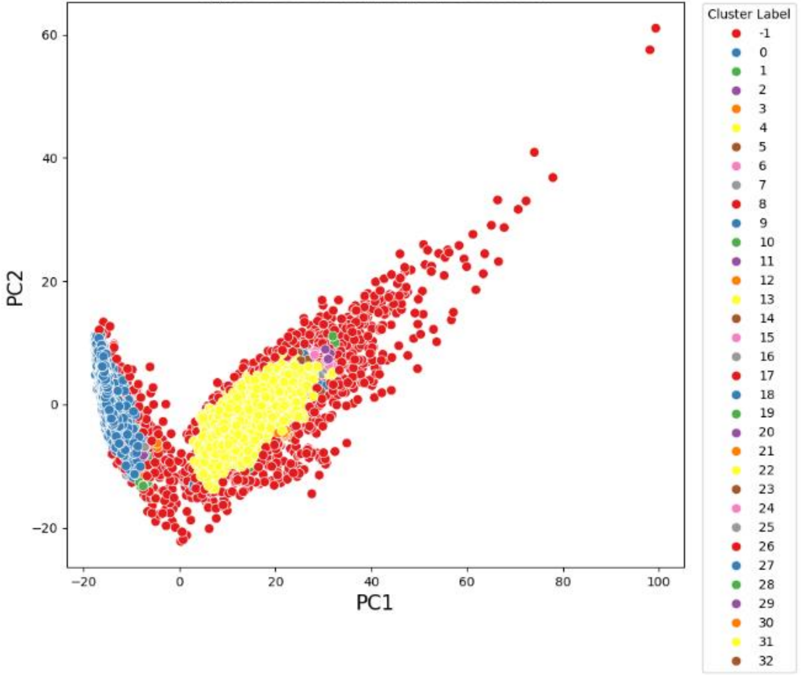
Fig. 1: DBSCAN Anomaly Detection

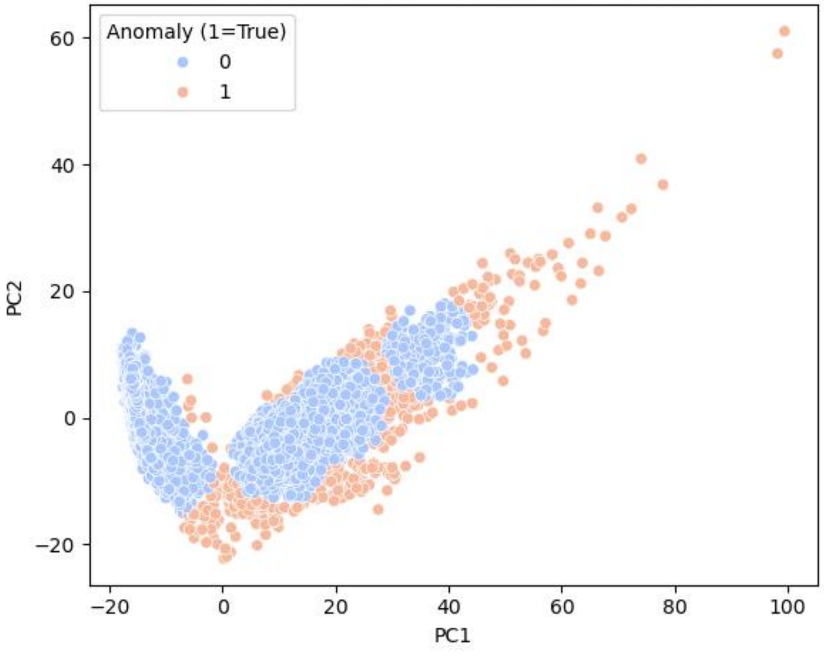
Fig. 3: K-Means Outlier Detection via Distance

- **DBSCAN** identifies both overall and local outlier instances while performing well on irregularly organized and noisy data distributions. The detection algorithm marked many outlier sparse points in border areas along with additional anomalies.
- **K-Means** assumes equal distance influence and spherical forms; it frequently misclassifies border points or outliers close to cluster edges. This unsupervised method does not provide reliable results or not robust to outliers when clusters differ in density pattern or shape distribution.

REFERENCE

- Zimek, Arthur & Schubert, Erich & Kröger, Peer. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining. 5. 363-387. 10.1002/sam.11161.
- Hodge, Victoria & Austin, Jim. (2013). A Survey of Outlier Detection Methodologies (Reprint)
- Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., & Parra, X. (2013). Human Activity Recognition Using Smartphones [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C54S4K.
- Scikit-learn Documentation: Clustering Methods.