Maching Learning - Mini Project 3

Thinakone Louangdy

February 2024

1 Introduction

Due to its numerous applications in the security, athletic performance, and health monitoring arenas, human activity recognition (HAR) has experienced a notable upsurge in popularity. Using a robust dataset gathered from thirty people' accelerometers and gyroscopes during a variety of activities, our study investigates HAR. This initiative aims to not only categorize the data but also to identify intricate patterns that can disclose details about the subtleties of human movement. Our goal is to interpret these patterns with the help of the reliable clustering methods K-Means and DBSCAN, contributing to the advanced knowledge of activity detection through machine learning algorithms.

2 Data Processing

We have a dataset that contains train and test set in text format alongside with features, features_info and activity labels. In train and test folder, there were also other files but we will only use **subject**, **X** and **y** together with activity_labels and features name to merge as one data frame for easier to implement our clustering task. We double check the data again for null and nan values before we proceed to the next step. We then use **Standardscaler** to normalized our data to standardized the range of continuous variables which is crucial for distance-based algorithms like K-Means and DBSCAN.

3 Modeling

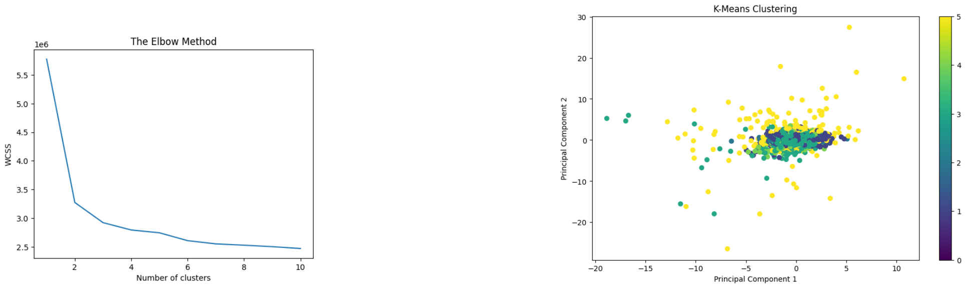
We were specified to use K-Means and DBSCAN to perform the clustering for this project. Therefore here is how we will construct our models:

- **K-Means**: The number of clusters was determined through methods like the elbow method, which identifies a point where the within-cluster sum of squares (WCSS) starts diminishing at a slower rate.
- **DBSCAN**: Parameters eps and min_samples were optimized through iterative testing via KDistanceGraph and tuning with silhouette score aiming to balance the discovery of meaningful clusters with the minimization of noise points and ensuring an effective balance in cluster identification.

3.1 Before Dimensionality Reduction

3.1.1 K-Means

We have selected clusters number based on the elbow method as mentioned above, we tested and try in between the range of the curve and the data points in the clusters exhibit varying degrees of spread and overlap, with some clusters tightly grouped indicating low within-cluster variance, and others more dispersed, suggesting higher within-cluster variance as we can see from figure 1.



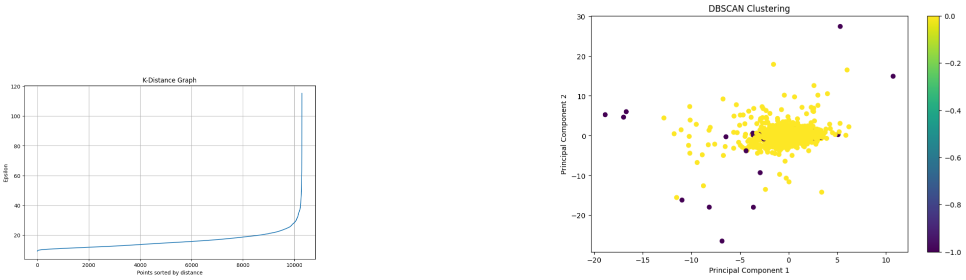
(a) Elbow Method

(b) Clustering

Figure 1: K-Means Clustering Performance

3.1.2 DBSCAN

We have selected clusters number based on the K-Distance Graph as mentioned above, we tested and try in between the range of the curve. As a result, there is a large, dense cluster in the center where points are closely packed, indicative of a high-density region. Surrounding this, there are scattered points, likely representing noise or outliers as identified by DBSCAN's algorithm, which are more isolated and spread out as we can observe from figure 2.



(a) K-Distance Graph

(b) Clustering

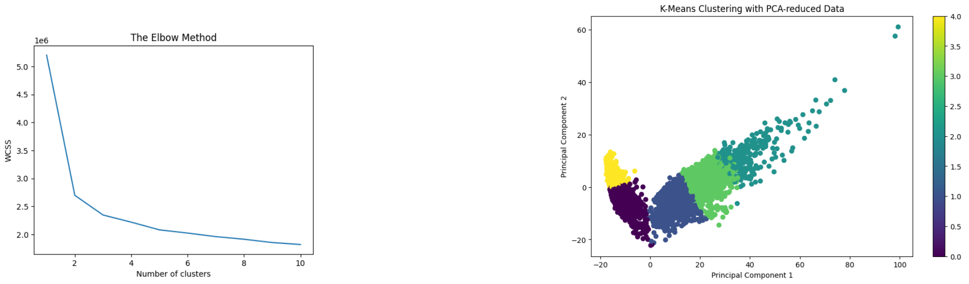
Figure 2: DBSCAN Clustering Performance

3.2 Principal Component Analysis (PCA)

With the requirement from this project, we firstly employed the principal component analysis (PCA) to our data for experiment with our clustering with K-Means and DBSCAN. We have applied PCA to our data while choosing to keep 90% of variance.

3.2.1 K-Means

The same method was applied as above, only switch to use feature after reduced dimensionality. As we can see from figure 3 after we perform a dimensionality reduction of our features, our cluster shows well-defined, dense clusters with points closely grouped together. Each cluster fans out from a central point, resembling a comet-like shape, indicating variation within clusters but clear separation between different clusters. There's a gradation in density from the cluster cores to their peripheries, which shows good clustering but with some spread within each cluster.



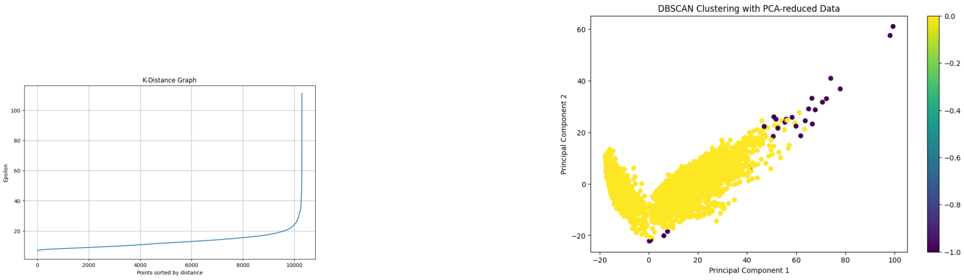
(a) Elbow Method with PCA

(b) Clustering with PCA

Figure 3: K-Means Clustering with PCA Performance

3.2.2 DBSCAN

The same method was applied as above, only switch to use feature after reduced dimensionality. We can observed from figure 4 that our DBSCAN Clusters shows one large, dense cluster with a broad spread of points along the primary axis of variation. There's a smaller, less dense purple cluster that appears detached from the main cluster which represent noise. Overall, we can still see some improvement after applied the dimensionality technique.



(a) K-Distance Graph with PCA

(b) Clustering with PCA

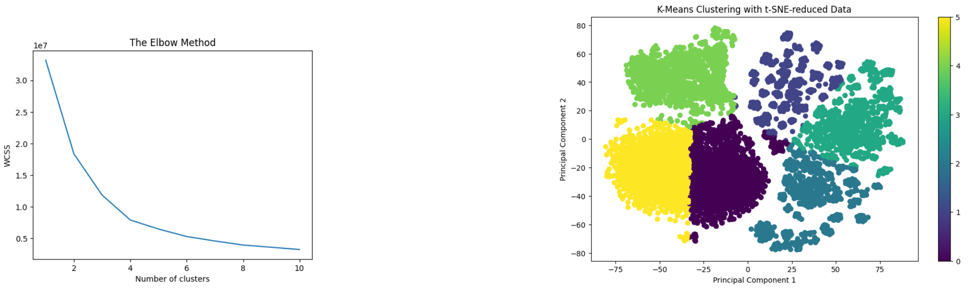
Figure 4: DBSCAN Clustering with PCA Performance

3.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

We have seen earlier after we applied PCA technique, there has been some improvement in both of out clustering but there were still some noise in out cluster. We will try t-Distributed Stochastic Neighbor Embedding (t-SNE) as a dimensionality reduction technique. While PCA is renowned for its ability to capture data variance in reduced dimensions, t-SNE excels in preserving the local structure of data and revealing the data's inherent clustering patterns at multiple scales.

3.3.1 K-Means

Same method applied, letś take a look at the result in figure 5. The clusters are sizable and distinct, each occupying a unique plot area with little to no overlap. The points within each cluster show a moderate spread, indicating a variation in the data yet maintaining clear cluster integrity. This separation suggests that the K-Means algorithm could identify cohesive groupings within the t-SNE-transformed space.



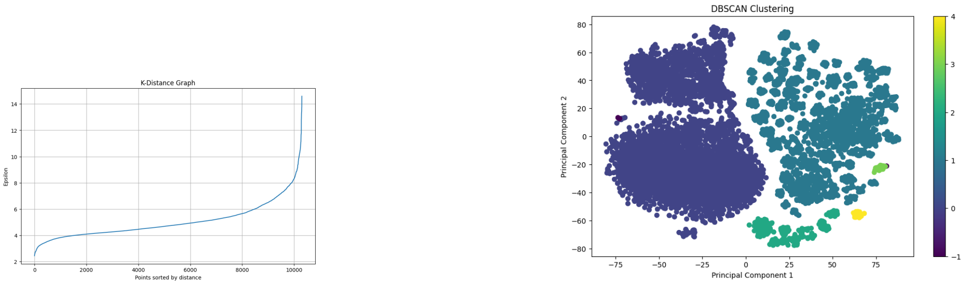
(a) Elbow Method with t-SNE

(b) Clustering with t-SNE

Figure 5: K-Means Clustering with t-SNE Performance

3.3.2 DBSCAN

We can observed from figure 6 that the spread of points within each main cluster is considerable, but the separation between the two large clusters is distinct, showcasing DBSCAN's ability to identify areas of high density and separate them from less dense regions, also there are several small clusters and isolated points, possibly representing outliers or noise.



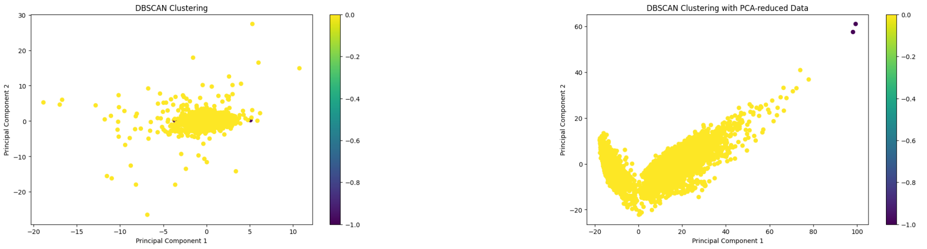
(a) K-Distance Graph with t-SNE

(b) Clustering with t-SNE

Figure 6: DBSCAN Clustering with t-SNE Performance

4 Silhouette Score for DBSCAN

We still see a strange behavior of DBSCAN Clusters from before dimensionality and after, without further due, we cannot relied only on K-Distance Graph. We want to experiment tuning the epsilons and min_samples value for our DBSCAN to decrease the spreading cluster we have seen above by calculating the Silhouette score which might improve cohesion and separation and handling noise of our cluster in DBSCAN.



(a) DBSCAN Fine-Tuned

(b) DBSCAN + PCA Fine-Tuned

Figure 7: DBSCAN Clustering Fine-Tuned

We chose to perform the tuning only for before and after dimensionality (only PCA) because there was already significant improvement on t-SNE for DBSCAN. As we witness from figure 7 has significantly improved our DBSCAN clusters. While before reducing the dimensionality, the spread within the main cluster seems tighter, and after reduced the dimension there was fewer noise and potentially yielding a more accurate representation of inherent data structure.

5 Conclusion

The human activity recognition dataset identified clusters corresponding to different activity patterns. Our study identified clusters corresponding to different activity patterns in the human activity recognition dataset. These patterns line up with the various movements that the participants wore sensors recorded. The principal scientific challenge is the calibration of K-Means and DBSCAN algorithms with particular parameter choices and dimensionality reduction obstacles. We overcome this by utilizing PCA and t-SNE for dimensionality reduction, the elbow approach, and silhouette scores to fine-tune the clustering parameters. Using these methods, the dimensionality of data was lessened, and cluster interpretability was enhanced. Other dimensionality reduction methods, such as UMAP, might be investigated in later research to improve clustering performance. Despite the challenges, the precise tuning of DBSCAN parameters and the strategic application of dimensionality reduction have provided a clearer, more structured view of the data, revealing the underlying patterns of human activities.