Unsupervised Analysis of Human Activity Recognition Data Using PCA and K-Means

Amirreza Dashti Genave

March 2025

Abstract

Human Activity Recognition (HAR) plays a crucial role in applications such as health monitoring. This work applies two unsupervised learning techniques, specifically Principal Component Analysis (PCA) and K-Means clustering, on the UCI HAR dataset [1]. This dataset contains inertial sensors data from a waist-mounted smartphone on 30 subjects performing daily living activities. We aim to reduce the dimensionality of the data while maintaining the meaningful information, and to identify the activity clusters using K-Means. Additionally, we visualize the inertia values to find the elbow point which helps determining the optimal number of clusters. Furthermore, we compare it to the original number of classes provided in the dataset.

1 Introduction

Human activity recognition has become an important subject in numerous fields such as medicine, technology, health, and diet. This activity primarily relies on wearable devices that use inertial sensors to measure movements. Accurate activity recognition can benefit our lives by enabling systems to understand our behaviors. The UCI Human Activity Recognition (HAR) dataset is a valuable resource for this task. In this work, we applied unsupervised learning methods, specifically Principal Component Analysis (PCA) and K-Means clustering, to analyze this dataset. We perform dimensionality reduction to retain meaningful data and apply K-Means clustering on the reduced version to capture and interpret the different types of activities. Moreover, we execute the K-Means algorithm with different numbers of clusters, determined by the dataset and the elbow method.

2 Dataset Description

The HAR dataset has been created in an experiment with 30 volunteers in the age range of 19 to 48. Subjects have been asked to perform six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist to capture 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The data has 561 variables divided into test and train sets which makes 10299 records in total, with no missing values and one label column.

However the dataset is labeled, unsupervised learning can help us visualize the data and find the hidden patterns. Moreover, the dimensionality reduction and clustering phases can help us find the natural groupings in the data.

3 Dimensionality Reduction with PCA

In this section, we standardize the input data using the `scale()` function to ensure that all variables are on the same scale. This step is needed before performing PCA, because PCA is sensitive to the scale of the input features. By scaling, we make sure that all variables contribute equally to the PCA.

After applying PCA on the data, an explained variance is calculated to determine how many components are enough to explain at least 90% of the data. This value is calculated as 65 components. Explained variance helps us understand how much information or variance each principal component captures.

To have a visualization of the data, we chose the first two principal components along with their true labels and created a temporary dataframe object. Figure 1 demonstrates the impact of PCA projection. As you can see, the data points have created two separate groups that can be identified visually. At this point, the class labels don't help much in identifying the groups.

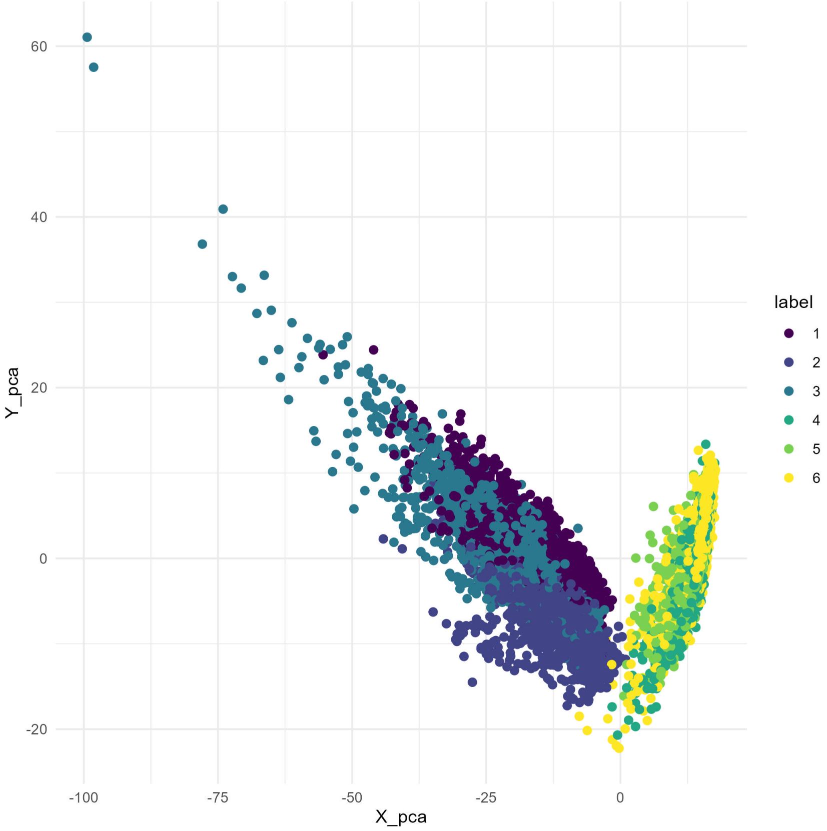

Figure 1: PCA projection showing two distinct groups in the data.

4 Finding the Optimal Clusters: The Elbow Method

The elbow method is a widely used technique to find the optimal number of clusters in a dataset. The idea is to calculate the within-cluster sum of squares (or inertia) for a candidate set of k. Here, k represents the number of clusters. By increasing the value of k, the inertia decreases. However, this decrease will slow down as we increase k. The objective is to find a point where the changes in inertia are no longer significant, and this point is called the ”elbow.” The elbow represents the optimal number of clusters derived from the elbow method.

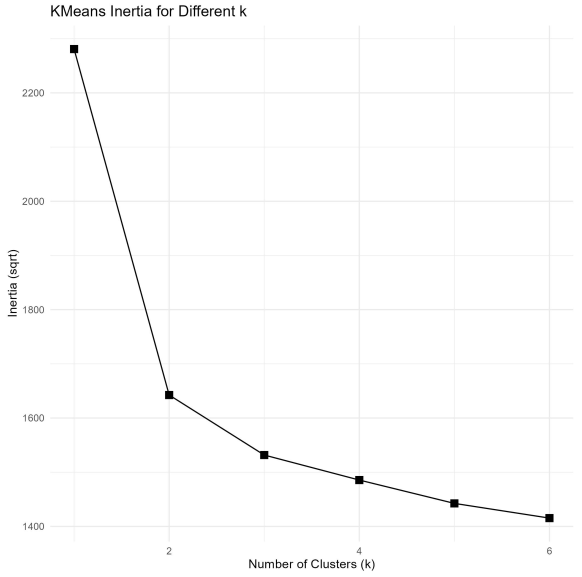

Figure 2: Elbow method for determining the optimal number of clusters

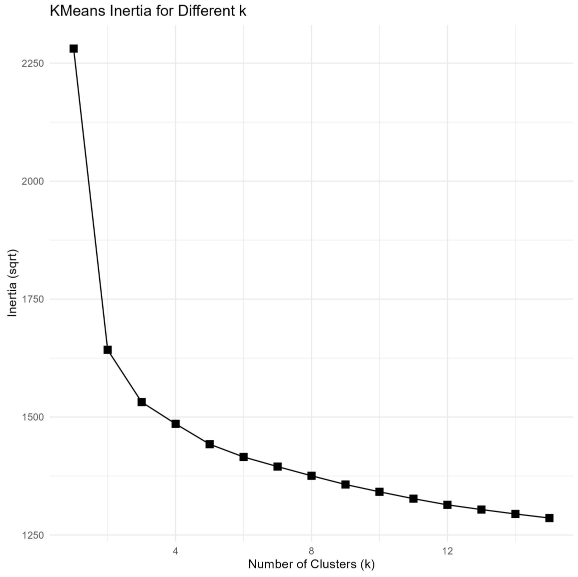

Figure 3: Elbow method for k = 1 to 15

5 K-Means Clustering on Reduced Data

K-Means clustering is a popular unsupervised learning algorithm that partitions the dataset into k clusters, where each data point belongs to the cluster with the nearest mean. It works by:

1. Choose k initial centroids.
2. Assign each data point to the closest centroid.
3. Calculate the centroids based on all data points assigned to that cluster.
4. Repeat until convergence.

Recalling the first step of this study, we performed the PCA dimensionality reduction to reduce the feature space. The purpose of that action was to speed up the K-Means algorithm and extract meaningful clusters. By applying K-Means with k set to the number of label classes, we can visualize how the original labels are distributed across clusters in the 2D space. Additionally, we compare these results with K-Means using $k = 2$. The plot for $k = 2$ shows two well-separated clusters of data points, highlighting the distinct grouping of the data. The yellow cluster in figure 4 includes labels 1 *WALKING*, 2 *WALKING_UPSTAIRS*, 3 *WALKING_DOWNSTAIRS*, and the purple cluster includes 4 *SITTING*, 5 *STANDING*, and 6 *LAYING* in the corresponding plot.

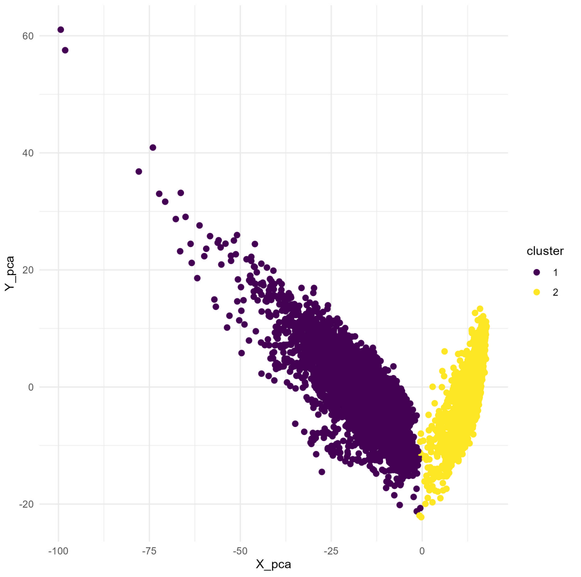

Figure 4: K-Means clustering with k=2

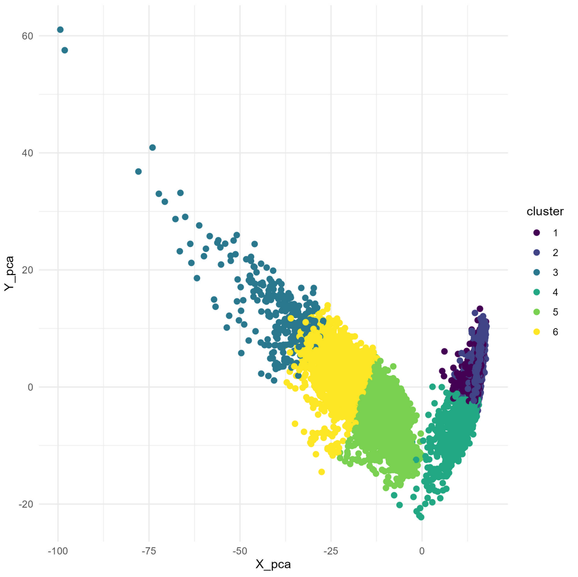

Figure 5: K-Means clustering with k=6

The insights from comparing figure 4 and 5, demonstrates $k = 2$ correctly identifies moving classes from stationary classes.

Code Availability

The code is available at: human activity recognition repository.

6 Conclusion

In this work, we applied unsupervised learning techniques, specifically Principal Component Analysis (PCA) and K-Means clustering, to analyze the UCI Human Activity Recognition (HAR) dataset. We retained the main features while simplifying the dataset using PCA, making it ready for further clustering algorithms.

Subsequently, we used the elbow method to find the optimal number of clusters for the K-Means algorithm. The plot demonstrated that the number of clusters significantly impacts the results.

Furthermore, using K-Means with different values of k, we were able to identify the natural structure of the data.

Overall, this work examines the unsupervised learning methods, and the effectiveness of dimensionality reduction in capturing the patterns in the data.

References

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “Human activity recognition using smartphones data set,” 2013. UCI Machine Learning Repository.