Machine Learning – Mini Project 3 – Otto Westerlund

Introduction

As technology is improving in both computational efficiency and cost efficiency, new application areas are emerging. Recently, Human Activity Recognition (HAR) has been of great research interest. HAR refers to the action of using computer technology to automatically recognize what a person is doing.

One HAR related experiment [1] had a sample of 30 people, on whom a Samsung smartphone were attached. The accelerometer and gyroscope of the phone were used to measure motion of the subjects during different activities, with the purpose of ultimately being able to automatically categorize what activity is being done based on the measured signals.

The purpose of this project is to inspect the collected data, and to see if patterns can be found in the measured signals.

Data set and processing

The data set was readily split into training and test data by the original research team, with a 70/30 ratio. In this project the training set was used for the experiment. The research team had done some pre-processing of the data. They had applied some filtering methods and normalized the data into a range of [-1, 1]. Indeed, as we load the data into a Pandas data frame and check for maximum and minimum values in the frame, -1 is the minimum and 1 the maximum value found. As such, the data wasn't further scaled or normalized in this project.

The data frame contained 7352 entries with 561 features or dimensions. The data frame was checked for missing values, none were found.

Pattern modelling

For the purpose of finding patterns, the data will be fed to a clustering algorithm. First, we wanted to visualize the data, which is impractical on 561-dimensional data. As such, dimensionality reduction using UMAP was done on the data, reducing the data to 3 dimensions. Figure 1 visualizes the data point distribution.

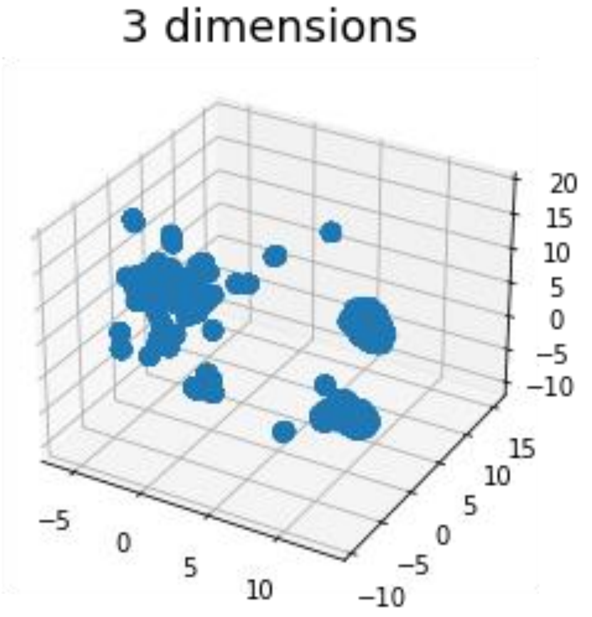

Figure 1. Data points in 3 dimensions.

It was unclear how reducing the dimensionality would affect the clustering process, because of that the clustering was ran on both the full-dimensional data and 3-dimensional data. For this project, DBSCAN algorithm was chosen. DBSCAN seemed appropriate since it could cluster the data into arbitrary shapes instead of just hyperspheres, meaning that it might detect patterns that e.g. k-means might miss.

DBSCAN requires the user to determine the values of two hyperparameters, epsilon and min_samples. To select epsilon, a nearest neighbor analysis was done on the data, to find out the distance distribution of the data points. Figures 2 and 3 show the distance distribution of the full-d and 3-d data. Epsilon was selected according to the slope of the graphs, and was ultimately set to 4.5 for the high-dimensional data and 0.2 for the low-dimensional data.

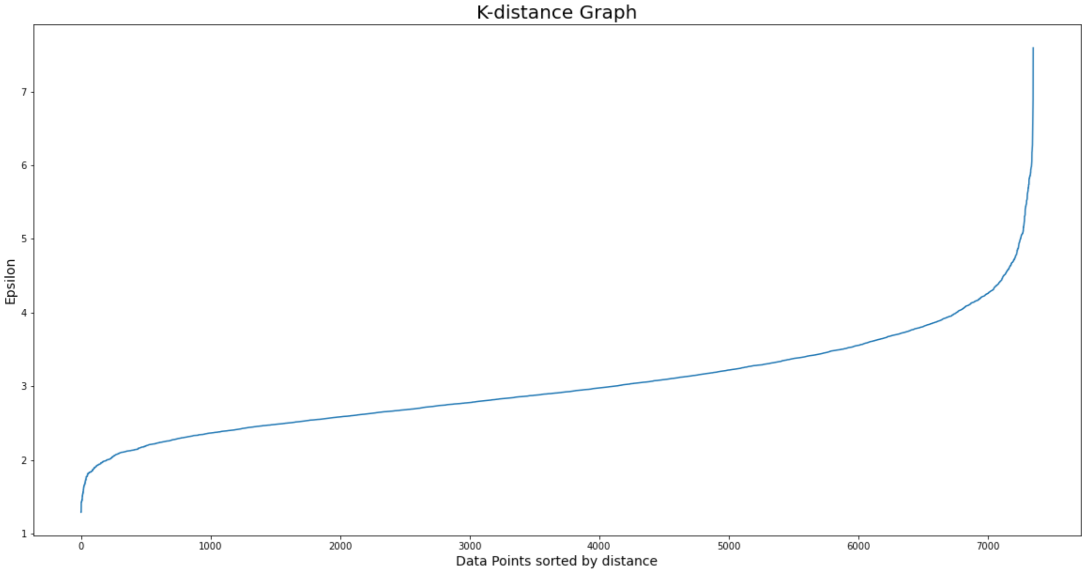

Figure 2. Distance distribution of high dimensional data.

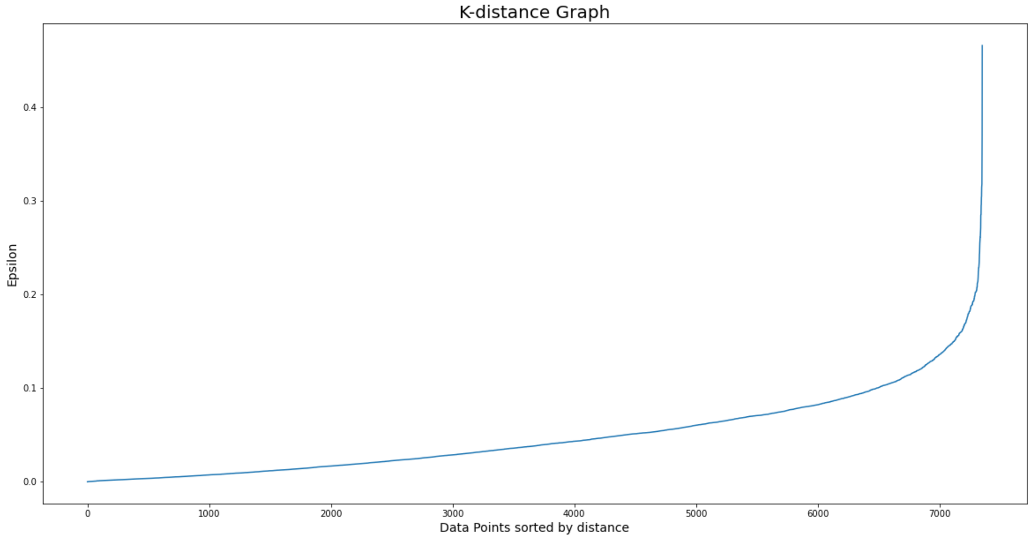

Figure 3. Distance distribution of low dimensional data.

For choosing min_samples, [2] suggested using ln(n). Based on this suggestion, min_samples was set to 9 for the high-dimensional data and to 10 for the low dimensional data. The low dimensional data got a similar value on the parameter, since setting it too low would include more noise in the clusters.

Results and Conclusions

After running the DBSCAN, it is clear that reducing the dimensionality plays a role in the outcome. In the high dimensional data DBSCAN found 3 clusters with 489 noise data points, while on the low dimensional data DBSCAN found 83 clusters with 484 points of noise. The conclusion is that due to the similar method of choosing epsilon and min_samples for both dimensionalities, DBSCAN labels similar data points as noise, but due to the different values of epsilon, it finds different amounts of clusters. The clusters can be seen in Figure 4.

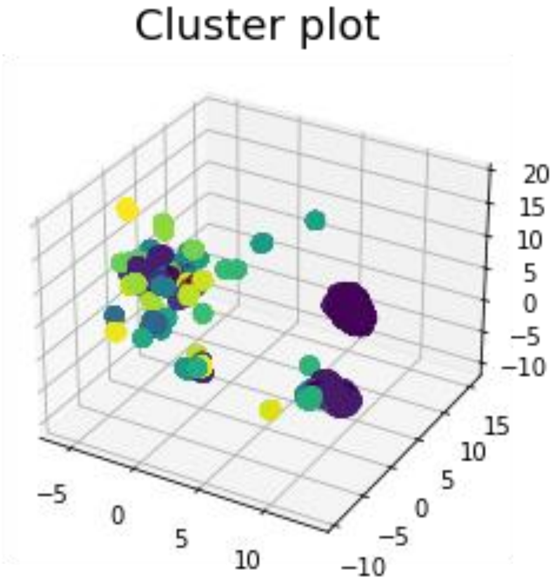

Figure 4. Data points grouped by DBSCAN.

One major point of interest was the computational cost difference in the dimensionalities. While the high dimensional data took 65 seconds for DBSCAN to cluster, the low dimensional finished in only 0.07 seconds.

Interpreting what the clusters represent presents some challenge. To understand that one has to consider what each dimension (of the high dimensional data) represents. The description of the features tells that the data represents the acceleration of the body and the gravity acceleration. These signals have been divided into X, Y and Z axes, and furthermore from these signals the research team derived Jerk signals, and the magnitude of each signal so far. Also, a Fast Fourier Transform has been applied to obtain frequency domain signals.

Considering all of that, one conclusion could be that the clusters represent the different actions that were measured by the research team (walking, walking upstairs/downstairs, sitting, standing, laying). Since DBSCAN found 3 clusters in the high dimensional data, the clusters could represent things like standing up (from a laying or sitting position), ascending or descending stairs, and walking. Looking at Figure 1, one can identify 3 distinct “blobs” of data points, which could perhaps represent the 3 distinct actions.

To test this hypothesis, the items in the training data set were mapped according to which cluster they belong and to what activity they correspond. The results were that 100% of cluster 0 corresponded to sitting, standing still, or laying down, and 100% of cluster 1 corresponded to walking, walking down the stairs, or walking up the stairs. A few items which corresponded to laying down were grouped into a separate cluster, cluster 2. The conclusion is that DBSCAN grouped the clusters according to how much motion was being measured in the signal.

References

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra and J. L, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," in *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013*, Bruges, Belgium, 24-26 April 2013.

[2] S. Tian, "Stackoverflow," 1 February 2018. [Online]. Available: https://stackoverflow.com/questions/12893492/choosing-eps-and-minpts-for-dbscan-r. [Accessed 23 February 2021].