Report 2: Detect activities

Arman Mohammadigilani, s301000,
ICT for Health attended in A.Y. 2022/23

January 17th 2023

1 Introduction

There are now more applications for inertial sensors than ever before because of advanced technologies, and one of them is the monitoring, identification, and categorization of human activities using sensors worn on the body [1]. The study of activity categorization is motivated by its potential applications in several fields. This report offers a comparative analysis of the various strategies to categorize the human activity that is being gathered through sensor data. This was accomplished by using the K-Means clustering technique and data analysis from multiple sensors attached to the subject's torso, right arm, left arm, right leg, and left leg [2]. The preprocessing of data and use of the K-Means method to categorize the activity being done have been detailed. In addition, the performance of the model and analysis results have been evaluated.

The public dataset that can be accessed from [3] is utilized in the following analysis.

2 Data analysis

The data were collected at a frequency of 25 Hz from 3-axes accelerometers, gyroscopes, and magnetometers, resulting in a total of 45 values for each sample. There are five spots on the subject's body where the sensors are implanted. The data were taken from 8 subjects that consist of 4 female and 4 male between the ages 20 and 30, each of which performed 19 different activities for a duration of 5 minutes. The signals were then divided into 5-second segments, resulting in a total of 480 segments for each activity.

The 19 activities that are classified using body-worn sensor units are listed in table 1. The analysis is carried for Subject 1.

3 Feature extraction

A discrete-time sequence has been obtained after capturing the signals as mentioned above. Minimum and maximum values, the mean, variance, skewness, kurtosis, autocorrelation

1	sitting	2	standing	3	lying on back
4	lying on right side	5	ascending stairs	6	descending stairs
7	standing in an elevator	8	moving in an elevator	9	walk.park
10	walk.4.fl	11	walk.4.15	12	run.8
13	exer.step	14	exer.train	15	cycl.hor
16	cycl.ver	17	rowing	18	jumping
19	playing basketball				

Table 1: List of activities

sequence, and the peaks of the Discrete Fourier Transform (DFT) of components at the relevant frequencies produce the very first collection of features.

3.1 Skewness

Skewed values for a given independent variable (feature) may invalidate model assumptions or obscure the significance of the feature, however this depends on the particular model in consideration. As a measure of the degree to which a collection of data differs from a perfectly symmetrical normal distribution (bell curve), skewness is a key statistic. Skewness is calculated as follows:

$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3} \tag{1}$$

where $\tilde{\mu}_3$ is skewness, N is the number of variables in the distribution, X_i is a random variable, $\bar{X}$ is a mean of the distribution, and σ is a standard deviation.

3.2 Kurtosis

How heavy or light the tails of the data are in comparison to a normal distribution is quantified by a statistic called kurtosis. What this means is that heavy tails, or outliers, are more common in data sets with a high kurtosis. Light tails, or a lack of outliers, characterize data sets with a low kurtosis. Kurtosis is calculated as follows:

$$Kurt = \frac{\mu_4}{\sigma^4} \tag{2}$$

where μ_4 is the fourth central moment, and σ is a standard deviation.

4 Preprocessing

Since each sensor unit on the torso, left hand, right hand, left leg, and right leg has 9 axes, the input segment must be collected over five seconds to collect 125 observations (5 x 25Hz) with 45 features. Then, a 1170x1 matrix is generated from the signals as follows:

The minimum, maximum, mean, skewness, and kurtosis for each of the 45 features measured over 125 samples, which adds up to a total of 225 features. The second step is to represent the maximum 5 peaks of the DFT applied on each of all features of the 5 units that consist of 225 features. Third, the frequency of the DFT's five peaks across the time series should be determined, yielding 225 features. The autocorrelation of the series is represented in the fourth stage by selecting 11 values out of 125 for each feature to represent it manually. This provides 495 features, 1170 in total for all steps.

Then, the subject ID and activity ID for that particular segment/text file are recorded with the normalized values in the range [0,1]. This will generate a 1140 x 1170 matrix for the subject 1.

5 K-means technique

In machine learning, K-means clustering is an unsupervised approach to finding groups of similar data points. Although there are other clustering algorithms available, K-means is an attempted technique that can be easily implemented. Data points are clustered into a specified number of clusters using this technique. The method determines which points belong to which clusters by constructing spherical clusters in which all points within a cluster are as near as feasible to the centroid (mean point) of that cluster while also being as distant as possible from the centroids of other clusters. The objective is to determine the optimal clustering of the data points such that the total of the distances between each point and its assigned cluster centroid is minimized. The purpose of this approach is to minimize the squared error function, which is defined as:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|\mathbf{x_i} - \mathbf{v_j}\|)^2 \tag{3}$$

where c is the number of cluster centers, c_i is the number of data points in i^{th} cluster, and $\|\mathbf{x_i} - \mathbf{v_i}\|$ is the Euclidean distance between $\mathbf{x_i}$ and $\mathbf{v_j}$.

6 Results

To organize the 19 distinct activities according to their characteristics, the K-Means classification method discussed in the previous section was utilized. There are 1140 feature vectors overall, all of which include features from the 5-second signal samples. As a result, instead of using slices in order to divide the data into training data and test data, only 20% of the feature vectors, after being generated using all four steps described in preprocessing section, are used in the training set, while the remaining 80% are used in the testing phase. Figure 1 shows the confusion matrix of the K-Means technique. It can be observed that first activity(sitting) and fourth activity(lying on right side) are the activities that are almost never confused. Confusion at the classification stage because of activities like seventeenth(rowing) and nineteenth(playing basketball) becomes inevitable.

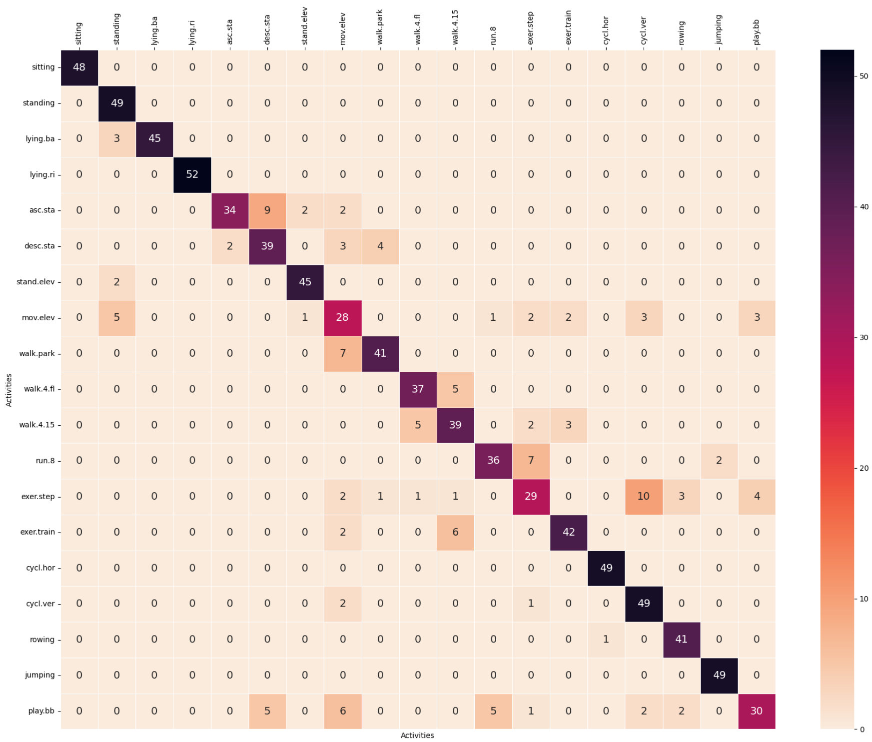

Figure 1: Confusion matrix for K-Means

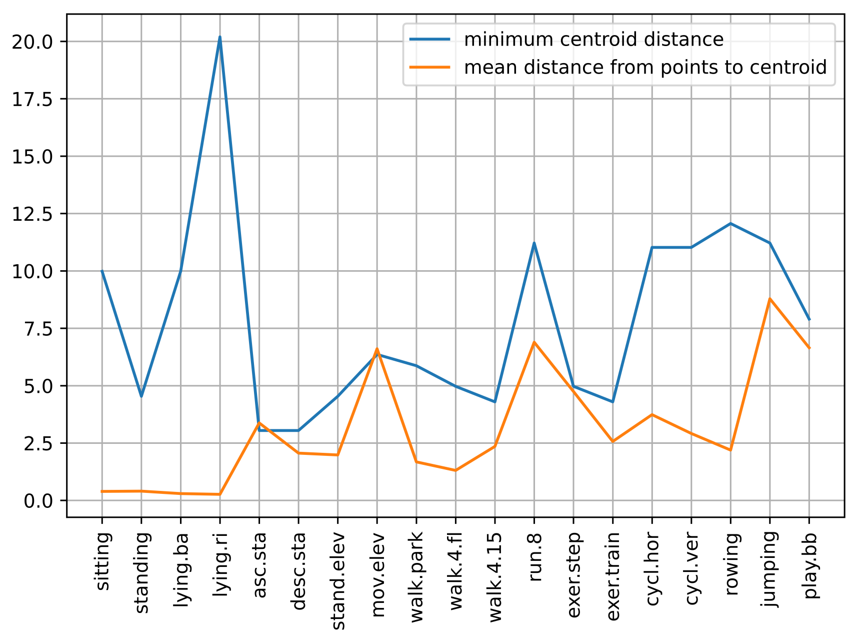

Figure 2: Cluster analysis

Figure 2 shows the Cluster analysis of the K-Means technique. After normalizing dataset and feature extraction, minimum centroid distance and mean distance from points to centroid has been calculated. Like 1, it shows that the blue line for the first and the fourth activity are higher than the other activities.

7 Conclusion

Based on the results of the analysis, it can be concluded that the K-Means clustering algorithm was able to classify the activity being performed by a subject based on sensor data collected from the subject's body. The accuracy of the classification was calculated to be 85.745%.

When visualizing the actual activity versus the predicted activity for each row of data, it can be can be seen that the model was generally able to accurately predict the activity being performed, with some exceptions. It's worth noting that the performance of the model may vary depending on the specific subject and the number of slices or the size of training and test sets. Additionally, the model's performance may be improved by incorporating additional preprocessing steps or by using a different clustering algorithm.

References

[1] W.Zijlstra, K.Aminian, Mobility assessment in older people: new possibilities and challenges, *Eur.J.Ageing4(1)(2007)3–12.*

[2] M.J.Mathie, B.G.Celler, N.H.Lovell, A.C.F.Coster, Classification of basic daily movements using a triaxial accelerometer, *Med.Biol.Eng.Comput.42(5) (2004) 679–687.*

[3] `https://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities`