Unsupervised Learning (Clustering) for Activity of Daily Living Recognition using Binary Sensor Dataset.

Priyanka Vyas
Department of Computer Science
Kent State University
Ohio, United States of America
pvyas2@kent.edu

Abstract—Clustering based approach is an unsupervised machine learning task. A cluster is a collection of datapoints that are similar to one another and this learning approach is the most fundamental approach to comprehend and learn information into logical categories or groups. Class labels, or category labels that refer to objects with historical identifiers, are not used in cluster analysis. Data clustering learning approach and classification learning approach are distinguished by the lack of category information. Clustering is exploratory in nature and aims to find structure in the data, this study focuses on unsupervised clustering learning.

Keywords—activity recognition, wireless sensor network, time series data, unsupervised learning, Kmeans algorithm.

I. INTRODUCTION

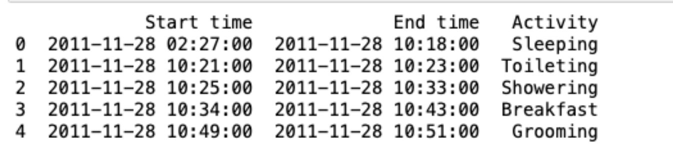
Recognizing Activities of Daily Living (ADLs) is an essential task for most of the monitoring systems. Clustering-based techniques have gained popularity as an ADL detection technique since binary sensor data is now more widely available. Using binary sensor datasets, this project provides a clustering-based method for ADL detection and the approach used is K-means technique. For unsupervised statistical learning, K-means clustering is a partitioning method. It differs slightly from aggregative techniques like hierarchical clustering. The goal of a partitioning strategy is to divide all of the data points into a certain number of clusters. A group of numerical variables are subjected to K-means analysis. This project is a study to predict the number of clusters and then make a prediction as to where their centers (also known as "centroids") are. In order to apply this clustering method repeatedly, the points are assigned to the nearest centroid before recalculating the centroids. In this study, Sensor data (features) will be used in further section to build the model and visualize the clusters. The ADL Label 'Activity' will be used to evaluate the overall accuracy and cluster performance.

II.DATA PROCESSING

A. Dataset introduction and pre-processing

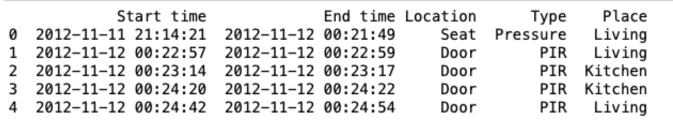
This Ordonez sensor events dataset comprises of two instances of the data, each corresponding to different user (A and B) and summing upto 35 days of fully labelled data. The features are the sensor events captured for corresponding sensor network.

OrdonezA sensor activity events:



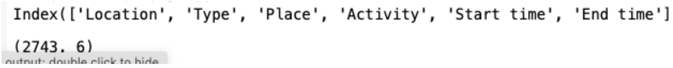
```
            Start time            End time   Activity
0  2011-11-28 02:27:00  2011-11-28 10:18:00   Sleeping
1  2011-11-28 10:21:00  2011-11-28 10:23:00  Toileting
2  2011-11-28 10:25:00  2011-11-28 10:33:00  Showering
3  2011-11-28 10:34:00  2011-11-28 10:43:00  Breakfast
4  2011-11-28 10:49:00  2011-11-28 10:51:00   Grooming
```

OrdonezB sensor events:



```
            Start time            End time Location      Type    Place
0  2012-11-11 21:14:21  2012-11-12 00:21:49     Seat  Pressure   Living
1  2012-11-12 00:22:57  2012-11-12 00:22:59     Door       PIR   Living
2  2012-11-12 00:23:14  2012-11-12 00:23:17     Door       PIR  Kitchen
3  2012-11-12 00:24:20  2012-11-12 00:24:22     Door       PIR  Kitchen
4  2012-11-12 00:24:42  2012-11-12 00:24:54     Door       PIR   Living
```

Below are the stages for preprocessing of data:
1. Data collection and cleaning: The data was collected from the sensor network and was available for access as text files. For this study, the data was imported from text to .csv using Microsoft excel. The data was formatted and cleaned for spaces and tabs using Microsoft excel data tool. After necessary cleaning and compilation, the final dataset is ready for with the respective input shape as below:



```
Index(['Location', 'Type', 'Place', 'Activity', 'Start time', 'End time']
(2743. 6)
output: double click to hide
```

Pandas library is used for storage and manipulation of tabular and time series data and it effectively performs the data cleaning, transformation and exploration. The dataset of two users is read and merge into one dataframe for further analysis and clustering. During the process of merging the dataframe, following issues were dealt like mismatched column names, missing values, duplicate values, mismatched data types. Due to these inconsistencies in data, the merging process continuously either produce incorrect result or empty dataframe for further analysis.

The missing values in the data frame were addressed by using fillna() method, in order to fill in missing values from other dataframe and the result is stored in 'Location', same approach is followed for Type, place Start time, End time and Activity columns. It is important to adress or get rid of missing values else it introduces bias and affects the analysis and accuracy.

2. Data Transformation and Feature Selection: In this study, few fields required data conversion like ‘Start time’ and ‘End time’ column to datetime format and location, type, place and activity into categories to use it into further analysis. This exercise also involved converting data into a format that is suitable for clustering and this also includes scaling the data to a common range or normalizing data to have a mean of 0 and standard deviation of 1.

Feature selection involves selecting a subset of features that are most relevant to the clustering task, in this study we are focused on features related to sensor events mainly Location, Place and type and it can also be done by using principal component analysis method. These features will be used in further section to build the model and visualize the clusters. The Label ‘Activity’ will be used to evaluate the overall accuracy and cluster performance.

III. Model Training and Evaluation

B. Model Design and Experimental Result

To build the cluster model first data standardization was done using StandardScaler, label encoder and one hot encode, but the model performs and showed optimum result through label encoder technique. The plot below relates to the cluster model where data is standardized using Standard Scalar and then KMeans algorithm is applied with 11 clusters and 42 as the random seed. The resulting clusters are plotted using principal component analysis (PCA) to reduce the dimensionality of the data to 2 dimensions. The resulting 2D scatter plot shows the clusters identified by KMeans. The ‘Activity’ label is dropped from the dataframe while model creation. Fig 1 and fig 2 shows the plot.

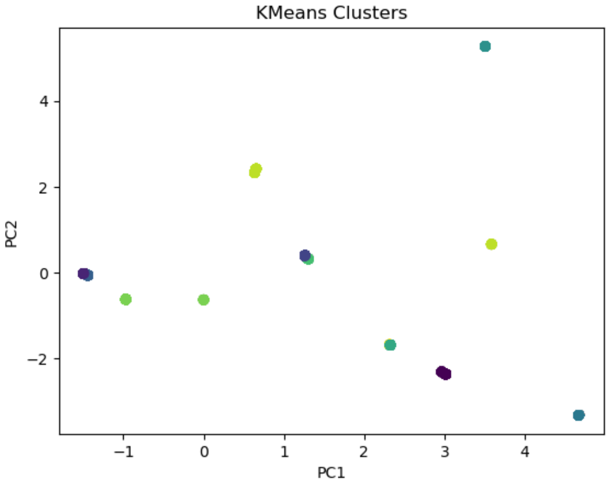


Fig1: Visualize Clusters (Standard scalar)

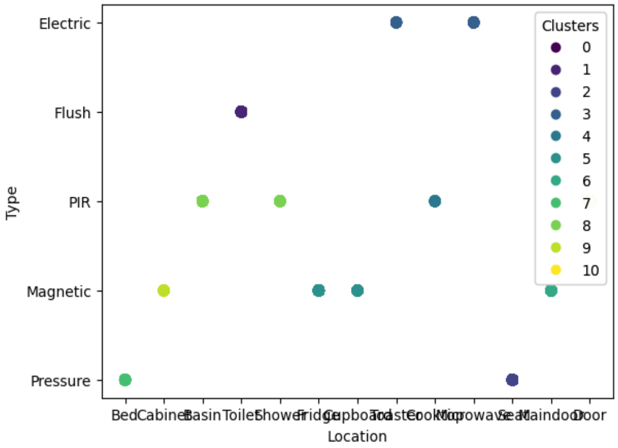


Fig2: Visualize Clusters (Standard scalar)

The second model performs k-modes clustering on the dataset using the kmodes library. The dataset is loaded from a CSV file and categorical variables are converted to numeric values using label encoding. The clustering is performed with k=11 and the resulting clusters are printed along with the cluster centroids. The overall accuracy and silhouette coefficient are calculated for the clustering. Fig 3 shows the plot with cluster centroids.

A confusion matrix, table 1 is also computed to evaluate the performance of the clustering. The resulting clusters are visualized using a scatterplot with different colors representing different clusters as shown below. Finally, homogeneity, completeness, and v-measure are calculated to evaluate the performance of the clustering.

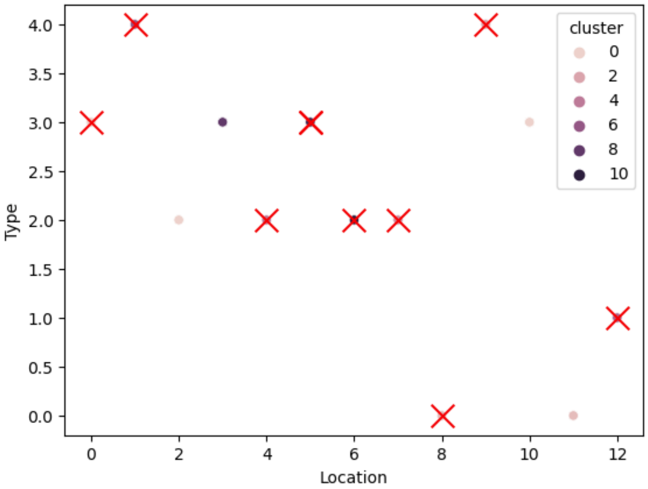
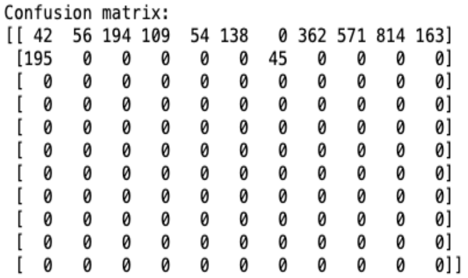


Fig2: Visualize Clusters (Label Encoder)



```
Confusion matrix:
[[ 42  56 194 109  54 138   0 362 571 814 163]
 [195   0   0   0   0   0  45   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]]
```

Table 1: Confusion. matrix

The confusion matrix shows the number of true positives, false positives, true negatives, and false negatives for each cluster. The rows of the matrix represent the true labels, while the columns represent the predicted labels. The confusion matrix provides more detailed information about how well the model is performing. Each row in the matrix represents the instances in a predicted class, while each column represents the instances in an actual class.

The numbers in the cells indicate the number of instances that were predicted to be in a particular class (row) and belonged to that class (column). In this case, the confusion matrix is showing that the model is predicting most of the instances to be in classes 9, 8, and 10, with many misclassifications. However, there are no instances in classes 2-7, which may indicate that the model is not properly trained or that the dataset is imbalanced.

C. *Model Performance Analysis*

Cluster model Performance	
Dataset	**Sensor Data**
Parameters	2743
No. of clusters	11
Homogeneity	0.092
Completeness	0.051
V-Measure	0.066
Accuracy	0.015
Silhouette Coefficient	0.90

In unsupervised learning, the accuracy evaluation is measured by comparing true label to predicted label. The above performance report explains that that the clustering algorithm did not perform well in terms of grouping similar data points together. Silhouette coefficient is a measure of how well-separated the clusters are in the clustering algorithm. A value of 1 indicates well-separated clusters, while a value of -1 indicates overlapping clusters. A value of 0 indicates that the clusters are very close together. In this case, the silhouette coefficient is quite high at 0.902, which suggests that the clusters are well-separated.

The accuracy score gives an overall measure of how well the clustering algorithm is performing. In this case, the overall accuracy is very low, only 1.5%. This suggests that the clustering algorithm is not performing well and needs to be improved. The Homogeneity measure ranges from 0 to 1, where a score of 1 means that all clusters contain only members of a single class. A score of 0.092 means that only 9.2% of the data points in each cluster belong to the same class. The Completeness measure ranges from 0 to 1, where a score of 1 means that all members of a class are in the same cluster. A score of 0.051 means that only 5.1% of the members of a given class were assigned to the same cluster. V-measure is the harmonic mean of homogeneity and completeness. It ranges from 0 to 1, where a score of 1 means that both homogeneity and completeness are perfect. A score of 0.066 means that the overall clustering performance is poor.

IV.EXPERIMENTAL ENVIRONMENT

Python, its libraries and TensorFlow is used to create, compile and train the neural network. To run the code in Jupyter, through anaconda navigator, an environment named 'TensorFlow' with necessary installed packages was first created to create the experimental environment for project. Apple M1 pro with GPU integrated on the same chip as CPU is the machine used to work on the project. Some limitations that were experienced while working on the data was to merge and clean the dataset. It interrupted the model training and also provided false or empty output.

CONCLUSION

This study proposes a clustering-based approach for ADL recognition using binary sensor datasets. The result demonstrates. The results demonstrates that through this approach with better quality and preprocessed data, ADL recognition is feasible. It is learned that even after data cleaning and preprocessing, during the merging significant missing value are introduced which in turn led to incomplete information about datapoint. The severity and pattern of missing values leads to such bias and incorrect results and also impacts the cluster performance.

Overall, the importance of carefully considering the presence of missing values is learned through this study. Unlike supervised learning, unsupervised learning doesn't have access to target variable, hence there are higher likely unlikely chances that missing values may problems like biased clustering or inaccurate estimates of data distribution. Few literatures Yuhan Jia, et al, and M. Cornacchia et al have proposed in their paper that supervised learning can be more effective when it comes to labeled data available for training and unsupervised learning is equally better as it reveals underlying hidden pattern and structure of data.

REFERENCES

- Ariza Colpas, P.; Vicario, E.; De-La-Hoz-Franco, E.; Pineres-Melo, M.; Oviedo-Carrascal, A.; Patara, F. Unsupervised Human Activity Recognition Using the Clustering Approach: A

Review. Sensors 2020, 20,2702.https://doi.org/10.3390/s20092702

- M. Cornacchia, K. Ozcan, Y. Zheng and S. Velipasalar, "A Survey on Activity Detection and Classification Using Wearable Sensors," in *IEEE Sensors Journal*, vol. 17, no. 2, pp. 386-403, 15 Jan.15, 2017, doi: 10.1109/JSEN.2016.2628346

- Yuhan Jia, Jianping Wu and Yiman Du, "Traffic speed prediction using deep learning method," *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, 2016, pp. 1217-1222, doi: 10.1109/ITSC.2016.7795712.