UNSUPERVISED LEARNING - FINAL ASSIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

In the following project we compared several clustering methods - K-Means, Fuzzy-C-Means, Gaussian Mixture Model, Agglomerative algorithm and Spectral clustering, and activate them on two different datasets in order to find the best clustering method. We used statistical tests the compare the quality of the clustering methods and Mutual Information method to measure the fit of clustering to an external classification, and in the end we found the outliers using DBSCAN (anomaly detection). Eventually we found out that the best clustering method is individual for each dataset and every dataset have different clustering method that suit him best.

1 INTRODUCTION

In this project we will deal with two large scale data sets with large dimensions – from visualizing the data in a figure (using algorithms such as PCA to reduce dimensions), to clustering the data with six different algorithms, computing the fit of the clustering to an external classification and using statistical test to find the best algorithm out of the six above. After all the above, we will use anomaly detection to find the outliers points.

The first dataset named 'HTRU2' - describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. The second dataset named 'MoCap Hand Postures' - a Vicon motion capture camera system was used to record 12 users performing 5 hand postures with markers attached to a left-handed glove.

2 METHODS

2.1 DATA HANDLING

For this mission I created a class named 'DataHandling'. This class gets the CSV file name (the dataset) and return the normalized and reduced dataset and tags in Pandas dataframe format (NOTE: I used only the first 40,000 rows of each dataset because of technical issues). The class also responsible to fill the missing values (such as '?' etc.) with the mean of their column. The dimension reduces performed by the PCA algorithm. The dimension of the produced dataset is 2, and the tags' dimension set to be 1.

2.2 DATA CLUSTERING AND VISUALIZATION

After we produce the dataset into pandas' dataframe format, we can use the produced dataset for clustering missions using the algorithms K-Means, Fuzzy - C- Means, GMM, Hierarchical clustering, Agglomerative clustering, Spectral clustering and DBSCAN. To use those algorithms, we must choose the right number of clusters to each dataset. How to do it? I used the elbow method and silhouette score test to determine the correct number of clusters. The rest of the settings set to be the default settings. For the task of visualizing the dataset I used the matplotlib library with default settings. You can find it under the class 'DataClustering' in file 'Functions.py' in the GitHub link.

2.3 STATISTICAL TESTS

The statistical tests used to evaluate the fit of clustering method to the external classification. I chose to use the T test for this assignment. First, to use the T test we must have a list of a few – In this case I

chose 13 samples - MI (Mutual Information) samples of the external classification of our algorithms. We get the MI scores from the algorithms' labels with the given tags (i.e., external classification). The T test return the P-value that measure the fit of the cluster when 1 value is best fit, and 0 value means bad fit.

2.4 ANOMALY DETECTION

I used the DBSCAN algorithm that is a clustering density-based (details in the sections above). I used eps = 0.3 because both datasets are very crowded so it will be hard to find the outliers with larger epsilon. For the min_samples value I took the default because of the same reason that the datasets are crowded.

3 RESULTS

3.1 CHOOSE CORRECT NUMBER OF CLUSTERS

In each dataset we used six different algorithms to cluster the data: K-Means, Fuzzy-C-Means, Gaussian Mixture Model, Agglomerative algorithm, Spectral clustering and DBSCAN. How to choose the correct number of clusters? To answer this question, we used several recommended methods for this assessment – the Elbow method and the Silhouette Score method as you can see in Figure 1:

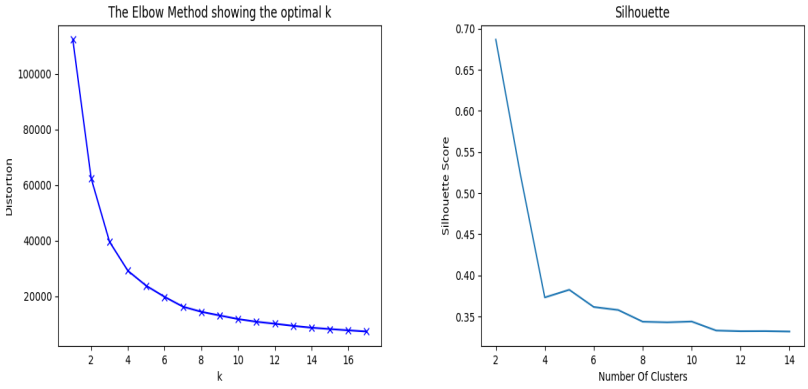

Figure 1: Elbow method and Silhouette score of HTRU2 dataset

According to Figure 1, the correct k (number of clusters) for dataset 1 (A.K.A 'HTRU2') is 2, so now we can run the algorithms with the correct k. Similarly, we can find that the correct number of clusters for dataset 2 (A.K.A 'MoCap Hand Postures') is 3.
NOTE: In the Elbow method we are looking for the bent in the graph, and in the Silhouette score we are looking for the largest value.

3.2 STATISTICAL TESTS

After running the algorithms, we would like to check the quality of the clustering with an external classification (In this case – the given classes / tags). For this assignment, we will use the Mutual Information method (specifically the Adjusted Mutual Information from the sklearn library in python). You can see the results in the Table 1 (for HTRU2) and Table 2 (for MoCap Hand Postures):

Table 1: Mutual Information score of each algorithm on HTRU2 and the given external classification

Clustering Method	Mutual Information Score
K-Means	0.596730
Fuzzy-C-Means	0.469273
Gaussian Mixture Model	0.412576
Agglomerative algorithm	0.573543
Spectral clustering	0.203190
DBSCAN	0.048842

Table 2: Mutual Information score of each algorithm on MoCap Hand Postures and the given external classification

Clustering Method	Mutual Information Score
K-Means	0.184540
Fuzzy-C-Means	0.185489
Gaussian Mixture Model	0.232822
Agglomerative algorithm	0.179339
Spectral clustering	0.144758
DBSCAN	0.002891

From the numbers above, we can infer that the clustering does not fit the external classification (1 value means 100% fit).

3.3 Best Clustering Method

Finally, we want to find the best clustering method out of the clustering methods we used in this section. We will do it using the T-test. We will run a few "battles" – in each one of them the T-test will get a list of few samples of the MI score of the clustering method. The better method (i.e., the better method according to the P-value reported from the T-test) will continue to the next "battle". The last method to stand - is the winner – i.e., the best clustering method for this dataset.
You can find the P-value report and the results in Table 3 and Table 4 below:

Table 3: "Battles" of HTRU2

First Algorithm	Second Algorithm	P-values
K-Means	Fuzzy C Means	2.8532101586945626e-14
Fuzzy C Means	Gaussian Mixture Model	8.52028941317485e-12
Gaussian Mixture Model	Agglomerative Clustering	0.9999999995885733
Gaussian Mixture Model	Spectral Clustering	0.9999999999699204
Gaussian Mixture Model	DBSCAN	1.0

Table 4: “Battles” of MoCap Hand Postures

First Algorithm	Second Algorithm	P-values
K-Means	Fuzzy C Means	1.0
K-Means	Gaussian Mixture Model	1.0
K-Means	Agglomerative Clustering	1.0
K-Means	Spectral Clustering	1.0
K-Means	DBSCAN	1.0

And these are the clusters of the winning algorithms - GMM on HRTU2 and K-Means on MoCap Hand Postures (Figure 2).

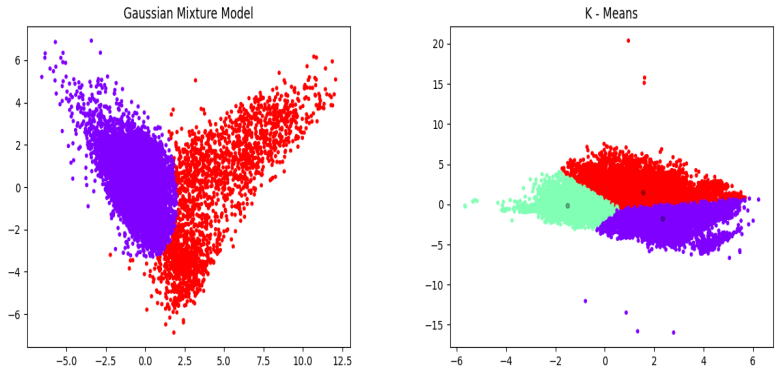

Figure 2: GMM on HTRU2 (left) and K-Means on MoCap Hand Postures (right) - the most accurate clustering for those datasets

3.4 Anomaly Detection

You can find the result of the DBSCAN algorithm (details under Methods section) right here (Figure 3 of HTRU2 and Figure 4 of MoCap Hand Postures):

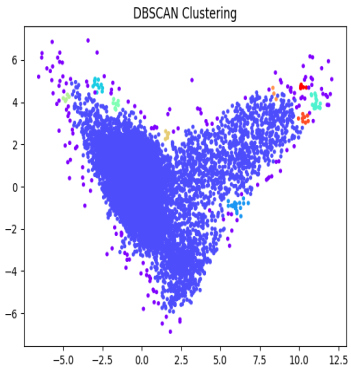

Figure 3: DBSCAN on HTRU2 - the cluster marked in purple, the rest are outliers

We can see that the outliers divided to a few clusters on the edge of the main cluster.

Same as the previous figure, the outliers divided to a few clusters on the edge of the main cluster.

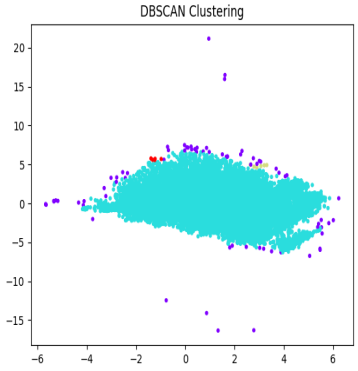

Figure 4: DBSCAN on MoCap Hand Postures - the cluster marked in turquoise, the rest are outliers

4 DISCUSSION

After reviewing the results, we can see that in each one of the datasets, the Mutual Information value of every one of the clustering methods is different – in the UTRU2 dataset most of the MI scores are around 0.5 i.e., the external classification has about 50% fit, while in the MoCap Hand Posters the MI score is around 0.2 i.e., the external classification has about 20% fit.
Another different between the two datasets is the best clustering method – GMM for HTRU2 and K-Means for MoCap Hand Posters. From this information we can infer that there is no such an algorithm that suits for all the datasets, because every dataset has its own features. i.e., we must pick the best method to our data. In addition, the P-values are almost perfect for both sides (aims to 1 or 0), so we can infer that the results are accurate and unequivocal.
Because of the difference in the clustering results between the different clustering methods, we need to check their accuracy and only after seeing the results, we will be able to choose the best clusters for our data. From the results of the Elbow method and the Silhouette score, we can see that they support the result of each other. In other words – both methods are good and return the same values (In our particular datasets).
Looking on the result of the anomaly detection, we can infer that both datasets are dense because most of the outliers located at the edge of the clusters.

5 REFERENCES

The code of this project - `https://github.com/Yoavpich/Unsupervised-Learning-Final-Assignment`

Hamerly, Greg, and Charles Elkan. "Learning the k in k-means." Advances in neural information processing systems. 2004

Reynolds, Douglas A. "Gaussian Mixture Models." Encyclopedia of biometrics 741 (2009).

Murtagh, Fionn, and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.1 (2012): 86-97.

Bezdek, James C., Robert Ehrlich, and William Full. "FCM: The fuzzy c-means clustering algorithm." Computers Geosciences 10.2-3 (1984): 191-203.

Ng, Andrew, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." Advances in neural information processing systems 14 (2001): 849-856.