UNSUPERVISED LEARNING - FINAL ASSIGNMENT

Yair Gat

January 2021

ABSTRACT

In this research, I cluster Two different data sets using five different clustering algorithms *(2.4)*, I showed and analyzed the results in purpose to find the best clustering algorithm for each data set. Additionally, I use an algorithm of anomaly detection *(2.10)* and show the impact of the anomaly points on the clustering quality . Finally, I figure out that, in HTRU2 Spectral is the best clustering algorithm both in internal and external terms, in AllUserData GMM has the best external fitment while Spectral provides the best internal quality. Moreover, I noticed that in most algorithms clustering without anomaly points are better both in HTRU2 and AllUserData .

1 INTRODUCTION

Clustering is a process that partitions a given data set into homogeneous groups based on given features. Each clustering algorithm has different assumptions on the data such that density, shape, distribution and etc. To provide optimal clustering, I cluster the data with six different clustering algorithms which are K-Means, GMM, FCM, Spectral, Hierarchial, and Dbscan. To check if there is any difference between the performance, I compare the performance by statistical tests. Furthermore, to identify data points that deviate from data sets normal behavior I use anomaly detection method.

2 METHODS

2.1 LOAD CSV

The data is stored in a CSV fotmatted file. To read and store the data I use the *NumPy* python package. The data contains numeric and non-numeric values. Unfortunately, we don't have tools to analyze non-numeric data. Therefore, I build a dictionary that holds the non-numeric values with numeric keys. In addition, Each data set has given external classifications labels, I stored that classifications as a 1-dimensional array using the dimension reduction method(if necessary).

2.2 DIMENSION REDUCTION

Each dataset has more than three dimensions, it is not possible to visualize such a features space. To overcome this problem I use the Principal Components Algorithm (PCA).

2.2.1 PCA

Principal Component Analysis (PCA), is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets. It does it by transforming a large set of variables into a smaller one that is still able to explain the variance of the data. The goal of PCA is to reduce the number of variables in a dataset, while preserving as much information as possible.

2.3 GET SAMPLE FROM DATA

Since the data sets are huge, I take samples from the data, the samples point that I take is determined by the random state. The function gets one parameter which is a random state and returns a data frame of samples.

2.4 CLUSTERING ALGORITHMS

Clustering is an unsupervised machine learning algorithm family. It helps in grouping data points. I use in the following six clustering algorithms:

- K- Means
- GMM- Gaussian Mixture Model
- Hierarchial clustering
- FCM - Fuzzy C Means
- DBSCAN

In each algorithm, I use three parameters: *i* - the random state of the sample, the default value of *i* is zero. *k*- The number of clusters. *plot* - If *plot* variable is True I plot the clustered data and else I don't, the default value of *plot* is True. *anomaly*- True if we want to remove anomaly points from the clustering and False otherwise.

2.5 SILHOUETTE

Silhouette score is also known as silhouette coefficient. It is a metric that calculates the goodness of a clustering technique. Its value range is $[-1, 1]$. Silhouette Score defined to be

$$\text{Silhouette} \triangleq \frac{(a-b)}{\max(a,b)} \tag{1}$$

where a is the average distance between each point within a cluster. b is the average distance between all clusters.
In the Silhouette method I get four parameters:
Clustering Algorithm - the silhouette will run over the data clustered by this clustering algorithm. Title - The title of the output plot. Plot - If the Plot is True I plot the score, else I don't.i - random state, the default value of i is zero.

2.6 STATISTICAL TEST

A statistical test provides a mechanism for making quantitative decisions about a process. Every clustering method which I used in my research is based on a random initialization. The randomness can affect the results, to minimize the effect and I had to use a statistical test.

2.7 T-TEST- TWO-SAMPLE ONE-TAILED T-TEST

The two-sample t-test, also known as the independent samples t-test, is a method used to test whether the unknown population means of two groups are equal or not. One-tailed test examine the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction.

I perform Two Sample One Tailed T-Test by the following way: suppose we have A, B lists of number so our H_0 is $\mathbb{E}[A] > \mathbb{E}[B]$. Then, $\overline{A} = mean(A), \overline{B} = mean(B)$

$$\text{p-value} = \begin{cases} \frac{p}{2} & \text{if } \overline{A} > \overline{B} \\ 1 - \frac{p}{2} & \text{otherwise} \end{cases} \tag{2}$$

If p-value is bigger than 0.05 then I got the H_0 and otherwise I rejected it.
This method gets one parameter which is *silhouette list* - A list that holds a silhouette value of five clustered data with the same algorithm but different random state.

2.8 OPTIMAL NUMBER OF CLUSTERS

All the cluster algorithms I use in my research has a random initialization. I sample 11000 points from the data. To provide the optimal number of clusters I run Silhouette 2.5 five times with different

11000 sample points each time with the function 2.3. Finally, I use two-sample one-tailed T-Test 2.6 to decide which is the optimal.
This function gets clustering algorithm as a parameter and returns the optimal number of clustering for the specific algorithm and data.

2.9 ADJUSTED MUTUAL INFORMATION

Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI), I use AMI to measure the dependency between the given classification and the clustered data.

$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{\overline{H(U), H(V)} - E(MI(U,V))} \tag{3}$$

where MI is Mutual Information
This function gets the following parameters:
- *method*- the clustering algorithm.
- *optimal number of clustering* - the optimal number of clustering specific to this data and algorithm.

2.10 ANOMALY DETECTION

Anomaly detection is a step in data mining that identifies data points that deviate from a dataset's normal behavior.

2.11 LOCAL OUTLIER FACTOR (LOF)

LOF compares the density of a given data point to its neighbors and determines whether that data is normal or anomalous.
This method gets a parameter plot that is True in case we interested to show the plot and returns 1/-1 ndarray, -1 if the point is an outlier, and 1 otherwise.

3 DATA SETS

3.1 DATA-SET 1: 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.The data describes a 3D motion of the glove by column x_i, y_i, z_i when i is in 0-11. I fill the empty features with the median of the feature's column. Additionally, the data contains two classification labels Class and User.

3.2 DATA-SET 2: HTRU2

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. Pulsars are a type of star, of considerable scientific interest. The data contains 17898 pulsar examples.

4 RESULTS

As I mentioned before I do a number of steps to find the best clustering algorithm and the impact of the anomaly points on the clustering quality. In the following tables, I will describe the result of each step that has led me to the final conclusion. Once, I will present the analysis on all data and second without the anomaly points, and accordingly, I will figure out the impact of the anomaly points on the results.

4.1 COMPLETE DATA-SETS

	Spectral	K-Means	Hierarchical	GMM	FCM
2	0.704	0.686	0.417	0.645	0.6513
3	0.608	0.528	0.445	0.485	0.348
4	0.602	0.376	0.46	0.389	0.364
5	0.58	0.38	0.42	0.35	0.347
6	0.57	0.36	0.416	0.3475	0.344
7	0.547	0.359	0.37	0.337	0.325

Table 1: HTRU2 - The average of silhouette score.

	Spectral	K-Means	Hierarchical	GMM	FCM
2	0.62347	0.453	0.401	0.358	0.457
3	0.457	0.4648	0.434	0.488	0.405
4	0.478	0.492	0.45	0.412	0.385
5	0.464	0.438	0.443	0.389	0.345
6	0.4585	0.435	0.422	0.389	0.3914
7	0.452	0.412	0.37	0.364	0.387

Table 2: All User Data - The average of silhouette score.

Both in table 1 and 2 we get a different score for each number of clusters which means there is an optimal number of clusters for each algorithm. To predict the optimal number I performed T-Test and the results in the following tables:
Lets defined $H_0(x, y)$ = x clusters is better then y clusters.

Spectral	K-Means	Hierarchical	GMM	FCM
$H_0(2,3) \sim 1$	$H_0(2,3) \sim 1$	$H_0(2,3) \sim 1$	$H_0(2,3) \sim 1$	$H_0(2,3) \sim 1$
$H_0(2,4) \sim 1$	$H_0(2,4) \sim 1$	$H_0(2,4) \sim 1$	$H_0(2,4) \sim 1$	$H_0(2,4) \sim 1$
$H_0(2,5) \sim 1$	$H_0(2,5) \sim 1$	$H_0(2,5) \sim 1$	$H_0(2,5) \sim 1$	$H_0(2,5) \sim 1$
$H_0(2,6) \sim 1$	$H_0(2,6) \sim 1$	$H_0(2,6) \sim 1$	$H_0(2,6) \sim 1$	$H_0(2,6) \sim 1$
$H_0(2,7) \sim 1$	$H_0(2,7) \sim 1$	$H_0(2,7) \sim 1$	$H_0(2,7) \sim 1$	$H_0(2,7) \sim 1$
2	2	2	2	2

Table 3: HTRU2 - P Value of T-Test

Spectral	K-Means	Hierarchical	GMM	FCM
$H_0(2,3) \sim 1$	$H_0(2,3) << 0.05$	$H_0(2,3) << 0.05$	$H_0(2,3) \sim 1$	$H_0(2,3) \sim 1$
$H_0(2,4) \sim 1$	$H_0(3,4) << 0.05$	$H_0(3,4) << 0.05$	$H_0(2,4) \sim 1$	$H_0(2,4) \sim 1$
$H_0(2,5) \sim 1$	$H_0(4,5) \sim 1$	$H_0(4,5) \sim 1$	$H_0(2,5) \sim 1$	$H_0(2,5) \sim 1$
$H_0(2,6) \sim 1$	$H_0(4,6) \sim 1$	$H_0(4,6) \sim 1$	$H_0(2,6) \sim 1$	$H_0(2,6) \sim 1$
$H_0(2,7) \sim 1$	$H_0(4,7) \sim 1$	$H_0(4,7) \sim 1$	$H_0(2,7) \sim 1$	$H_0(2,7) \sim 1$
2	4	4	2	2

Table 4: All User Data - P-Value of T-Test

By looking at tables 3 and 4 I can figure out from the last row that the optimal number of clusters for each algorithm. In the following Tables I measure the external quality of the classification by AMI (2.9):

height	Spectral	K-Means	Hierarchical	GMM	FCM
HTRU2	0.0267	0.0237	0.02	0.0175	0.021
All User Data	0.0421	0.2469	0.242	0.27	0.1785

Table 5: Adjusted Mutual Information

4.2 Discussion

As I have already mentioned, Silhouette is a measurement of internal quality clustering, AMI is a measurement of external quality. So to checking which algorithm has the best internal quality, I throwback to tables 1 and 2, and the algorithm that has the maximum Silhouette score is the one with the best internal quality. In addition to checking which algorithm has the best external quality I throwback to table 5 and check which algorithm has received the maximum AMI value. For HTRU2 Spectral is the best clustering algorithm in internal and external terms. For AllUserData GMM has the best external fitment while Spectral provides the best internal quality.

4.3 Clustering quality - anomaly points removed

To analyze the impact of anomaly points on the clustering quality I do the same process that I described in the section above with anomaly points removed.

Lets defined p to be the optimal number of clusters and Silhouette(p) to be the silhouette average of the optimal number of clusters

	Spectral	K-Means	Hierarchical	GMM	FCM
p	2	2	2	2	2
Silhouett(p)	0.703	0.686	0.716	0.646	0.651

Table 6: HTRU2- Silhouette average score of the optimal number of clusters

	Spectral	K-Means	Hierarchical	GMM	FCM
p	2	4	2	6	4
Silhouett(p)	0.47	0.5	0.46	0.4	457

Table 7: All Users- Silhouette average score of the optimal number of clusters

Tables six and seven described the internal quality of the algorithms. Surprisingly, there is no significant difference between the internal quality when we cluster the complete data and the data without anomaly points but if the difference exists the internal quality is better without the anomaly points.

height	Spectral	K-Means	Hierarchical	GMM	FCM
HTRU2	0.0274	0.0237	0.02	0.0173	0.020
All User Data	0.0423	0.2469	0.242	0.27	0.1785

Table 8: Adjusted Mutual Information- External Quality

In tables 4.1 and 4.3 we also can see a small impact of the anomaly points on the clustering quality.

5 SUMMARY

Firstly, I tried to answer the questions of "what is the best clustering algorithm" as we can see above this question is a little complicated because it's deepened on a lot of parameters like the data shape, data distribution, etc. Secondly, I tried to find the impact of anomaly points on the clustering quality, and I find out that the clustering quality is a little better when removing the anomaly points from the data.

REFERENCES

[1] Greater NoidaMichel Goossens, Frank Mittelbach, and Alexander Samarin. *Enhancing K means by unsupervised learning using PSO algorithmn*. A. Gupta, V. Pattanaik and M. Singh, "Enhancing K means by unsupervised learning using PSO algorithm," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 228-233, doi: 10.1109/CCAA.2017.8229805.

[2] Gaussian Mixture Models- Douglas Reynolds `http://leap.ee.iisc.ac.in/sriram/teaching/MLSP_16/ref`

[3] Dr.Ira Cohen
`https://www.anodot.com/blog/what-is-anomaly-detection/`