UNSUPERVISED LEARNING - MIDDLE ASSIGNMENT

Yair Gat

January 2021

ABSTRACT

In this research, I cluster three different data sets using five different clustering algorithms. The algorithms I use are *Gaussian Mixture Model, K-Means, Fuzzy-C-Means, Spectral Clustering,* and *Hierarchical Clustering.* I use the *Silhouette Score* and adjusted mutual information metrics to evaluate the model performance. Finally, to verify the significance of the results, I use the T-test.

1 INTRODUCTION

Get three long data sets which are: *Online Shoppers, Diabetic-Data, and E-Shop Clothing 2008*, and the research goal is to provide the optimal clustering for each data and figure which algorithm is the best. To do so, at first, I prepare the data for working by changing the non-numeric values to numeric. I use *PCA* to reduce the dimension of the data for visualizing and efficiently analyzing. In the entire research and each data-set, I worked with 11,000 random samples for making sure that . I cluster the data with five different algorithms, and for checking if there is a difference between the algorithms I used a statistical test and finally, I provide the optimal cluster for each data and deduced which algorithm is the best.

2 METHODS

2.1 UPDATE CSV

I get the data as CSV file, read and store the data using the *numpy* library. The data contains numeric and non-numeric values, unfortunately, we don't have tools to analyzing with non-numeric data, therefore we build a dictionary that holds these values with numeric keys and change the non-numeric to be the / fits? numeric key. Each data set has given external classifications column and I stored the classification as a 1-dimensional array using the dimension reduction method(if necessary).

2.2 DIMENSION REDUCTION

All our data is more than 3 dimensions, analyzing data with a high dimension is not especially intuitive, very clumsy, and impossible to visualize. Therefore I reduced the data dimension by using Principal Components Algorithm which also known as *PCA*, a dimension reduction algorithm. I stored the data after dimension reduction using *pandas* library with DataFrame.

2.2.1 PCA

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.The goal of PCA is to reduce the number of variables of a data set, while preserving as much information as possible.

2.3 CLUSTERING ALGORITHMS

Clustering is a Machine Learning technique that involves the grouping of data points. All the method I used enables an unsupervised-learning.

2.3.1 K-MEANS

K-Means is a hard clustering and very well-known algorithm. The main element of the algorithm works by a two-step process called expectation-maximization, one of many EM clustering algorithm. K-MEANS can find just circular clusters.
The Algorithm:

Specify the number k of clusters to assign.
Randomly initialize k centers.
while *The center does not change anymore.* **do**
Expectation: Assign each point to its closest center
Maximization: compute the new mean- the center of each cluster.

2.3.2 GMM- GAUSSIAN MIXTURE MODEL

Gaussian Mixture Model is a soft clustering algorithm. The reason is- every point has a 0-1 chance related to each cluster and the chosen cluster will be the one with the highest chance. GMM is also an Expectation-Maximization algorithm.
The Algorithm (In Shortly):

The Algorithm:

Start with random Gaussian parameters (θ)
while *The center does not change anymore.* **do**
: Compute $p(z_i = k|x_i, \theta)$. Check if sample i look like it came from cluster k.
Maximization: Update the Gaussian parameters (θ) to fit points assigned to them.

2.3.3 FCM - FUZZY C MEANS

Fuzzy C Means is a soft clustering algorithm which includes an element from K-MEANS and GMM and It is also an EM algorithm. The Algorithm:
1:Define target function L =

2.3.4 SPECTRAL CLUSTERING

Spectral clustering is an algorithm clustering based on graph theory, where the approach is used to identify communities of nodes in a graph based on the edges connecting them. The method is flexible and allows us to cluster non-graph data as well.

2.3.5 HIERARCHICAL CLUSTERING

Hierarchical clustering determines cluster assignments by building a hierarchy. It's a bottom-up approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
The Algorithm:

Make each data point a single point cluster)
Take the two closest data points and make them one cluster.
while *We have more than one cluster* **do**
:Take the two closest clusters and make them one cluster.

2.4 SILHOUETTE

Silhouette Score, also known as silhouette Coefficient is a metric used to calculate the goodness of a clustering technique. Its value range from -1 to 1.
Silhouette Score defined to be: $\frac{(a-b)}{max(a,b)}$ where: a = the average distance between each point within a cluster. b = the average distance between all clusters.

2.5 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 initialized and it can affect the results, so to minimize the effect and provide the best performance I had to use a statistical test.

2.6 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 tests 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:
Supposed we have A, B lists of number so our H_0 is E[A] $>$ E[B]

$p - value$ = $\begin{cases} \frac{p}{2} & \text{if } mean(A) > mean(B) \\ 1 - \frac{p}{2} & \text{otherwise} \end{cases}$
If $p - value$ bigger than 0.05 I got the H_0 and else I rejected it.

2.7 GET OPTIMAL NUMBER OF CLUSTERING

All the cluster algorithms I used in my research have random initialize and I worked with random samples, so to provide the most optimal number of clustering I ran Silhouette*(2.4)* five times with different sample points each time, and finally, I used T-Test- Two-Sample One-Tailed T-Test*(2.6)* to decide which Silhouette Score gives the most optimal results.

2.8 ADJUSTED MUTUAL INFORMATION

Adjusted Mutual Information (AMI) is an adjustment Mutual Information (MI), I consumed AMI to measure the dependency between the given classification and the clustered data.
$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{avg(H(U),H(V)) - E(MI(U,V))}$ When MI is Mutual Information.

3 RESULTS

3.1 DATA-SET 1: ONLINE SHOPPERS INTENTION

The first data-set consists information about users in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. It contains 12,330 instances and 18 features.

3.1.1 SILHOUETTE VISUALIZING

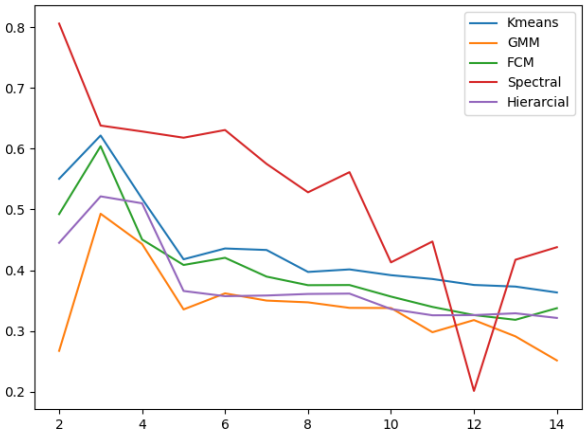

Figure 1: The maximum of Silhouette score for K-Means, FCM, Hierarchial and GMM received with 3 clusters and for Spectral 2 clusters. The method with the highest Silhouette score is Spectral.

3.1.2 CLUSTERING VISUALIZATION

The method I have described in *2.7* established the progress from *3.2*.

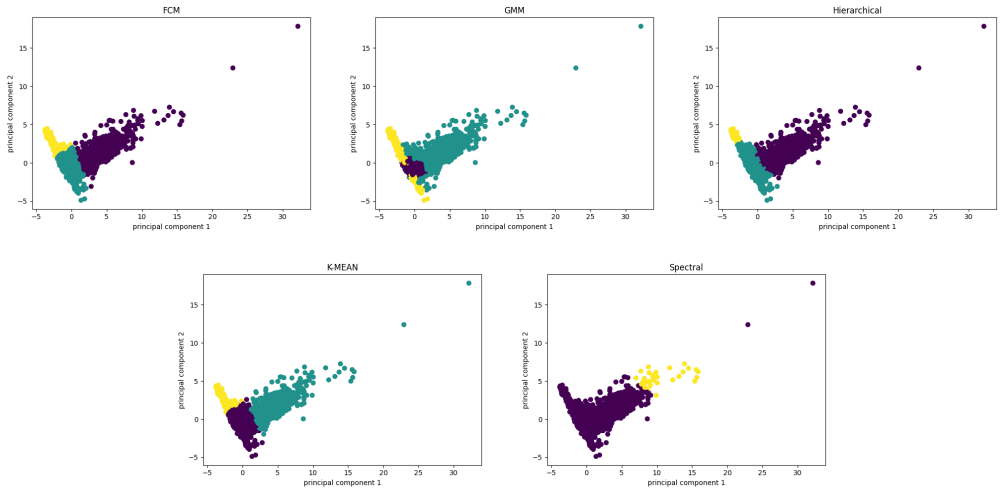

Figure 2: Data Set 1: Optimal clustering

3.1.3 EXTERNAL QUALITY OF CLUSTERS

Data 1- Adjusted Mutal Information Of the given classification and the clustered data

Algorithm	Optimal Number Of Clustering	AMI
Spectral	2	0.00182358
K-Means	3	0.03213964
GMM	3	0.031893825
Hierarchical	3	0.03193426
FCM	3	0.032864301

Table 1: As we can figure out all the algorithms give almost the same AMI but the one with the highest is FCM, so for data-set 1, the FCM algorithm has the highest fitment to the external classification.

3.2 DATA-SET 2: DIABETIC DATA

The sconed data-set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It contains 10,000 instances and 55 attributes.

3.2.1 SILHOUETTE VISUALIZING

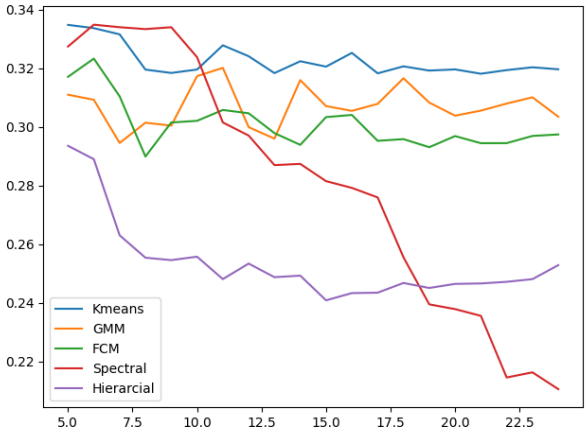

Figure 3: The maximum of Silhouette score for FCM and Spectral recived with 7 clustering, Hierarchial, GMM and K-Means received maximum with 5 clusters. The method with the highest Silhouette score is K-Means

3.2.2 CLUSTERING VISUALIZATION

For Hierarchical, FCM, Spectral, and GMM, the method I have described int *2.7* estimated the results from *3.6* but, method *2.7* determined that the optimal number of clustering for K-Means is 12.

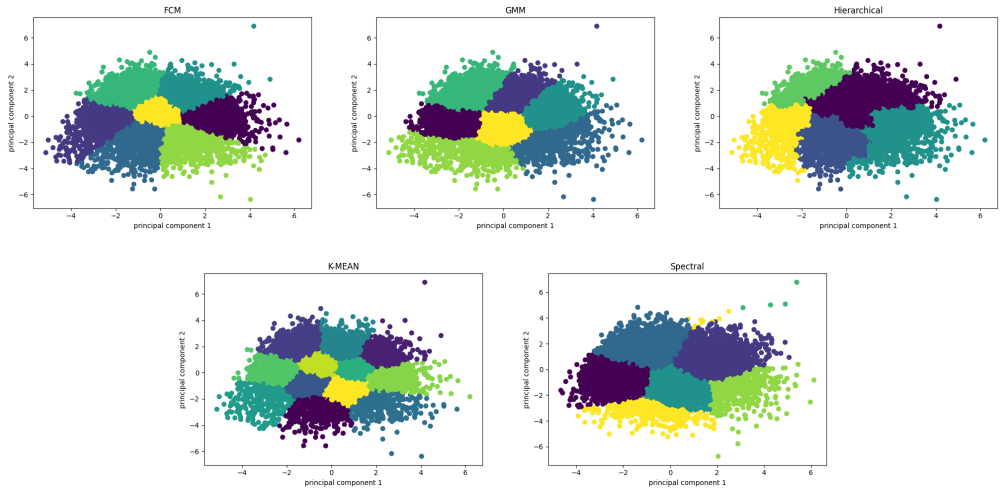

Figure 4: Data Set 2: Optimal clustering

3.2.3 External Quality of Clusters

Data 2- Adjusted Mutal Information Of the given classification and the clustered data

Algorithm	Optimal Number Of Clustering	AMI
Spectral	7	-0.0005323
K-Means	12	0.000331625
GMM	7	5.13E-04
Hierarchical	5	0.000121085
FCM	7	0.000464138

Table 2: All the algorithms give a very low value, The reason is in section *4*. The cluster with the highest AMI is FCM, therefore the most fitment clustering algorithm to the given classification is FCM.

3.3 Data Set 3: E-Shop Clothing 2008

The third data-set contains information on clickstream from an online store offering clothing for pregnant women. Data are from the five months of 2008 and include, among others, product category, location of the photo on the page, country of origin of the IP
address, and product price in US dollars. The data has 165474 instances and 14 attributes.

3.3.1 SILHOUETTE VISUALIZING

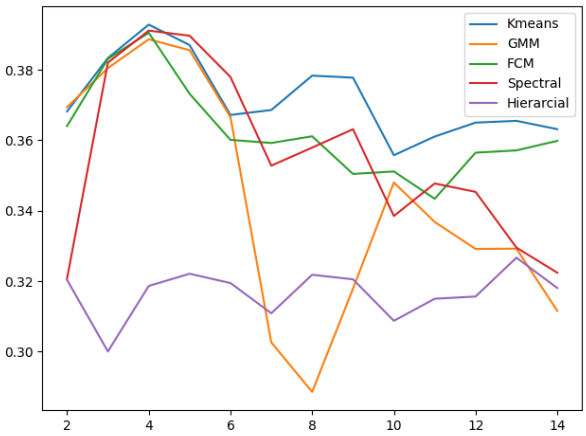

Figure 5: The maximum of Silhouette score for K-Means, GMM, FCM and Spectral received with 4 clusters but for Hierarchical Clustering received with 14. The cluster with highest Silhouette Score is K-Means.

3.3.2 CLUSTERING VISUALIZATION

Also here method *2.7* established the optimal number of clustering for K-Means, GMM, Spectral, and FCM but by *2.7* the optimal number of clustering for Hierarchical Clustering is 14.

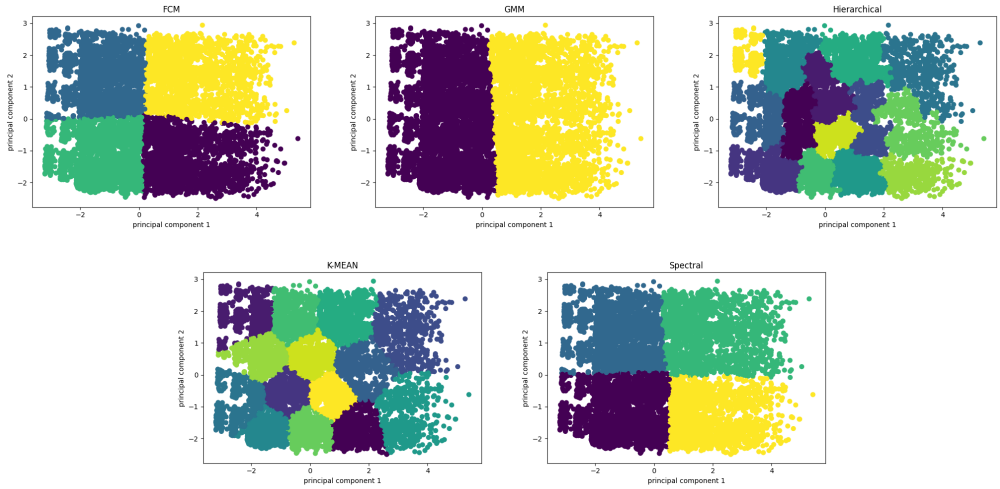

Figure 6: Data Set 3: Optimal clustering

3.3.3 EXTERNAL QUALITY OF CLUSTERS

Data 3- Adjusted Mutal Information Of the given classification and the clustered data

Algorithm	Optimal Number Of Clustering	AMI
Spectral	4	0.012642584
K-Means	4	0.012798789
GMM	2	0.011931646
Hierarchical	14	0.012980438
FCM	4	0.012642584

Table 3: The cluster with the highest AMI is Hierarchical, therefore the most fitment clustering algorithm to the given classification is Hierarchical.

4 DISCUSSION

In my research, I tried to answer the question of which clustering algorithm is the best one. First, as we saw this question is pretty general, I showed that the algorithm with the best internal fitment is not the one with the best external fitment. Second, as I mentioned before the adjusted mutual information is very low in all the algorithms because the external classification is not correlated to the internal data. As we can see in the table below *K-Means* gives the best clustering in internal-term for two out of three data-sets, while *FCM* gives the best external fitment also for two out of three, so we can figure out that *K-Means* gives the best internal fitment in the most of cases and *FCM* gives the best external fitment for the most of cases.

Data-Set	Highest Silhouette Score Algorithm	Highest AMI Algorithm
Data 1	Spectral	FCM
Data 2	K-Means	FCM
Data 3	K-Means	Hierarchical

Table 4: The best algorithm for each data-set, internal and external.

REFERENCES

[1] Gaussian Mixture Models- Douglas Reynolds
http://leap.ee.iisc.ac.in/sriram/teaching/MLSP$_1$6/$refs/GMM_Tutorial_Reynolds.pdf$

[2] Greg Hamerly, Charles Elkan - Learning the k in k-means
https://papers.nips.cc/paper/2003/file/234833147b97bb6aed53a8f4f1c7a7d8-Paper.pdf

[3] UCLA - The difference between one-tailed and two-tailed tests.
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-difference

[4] Gaussian Mixture Models- Douglas Reynolds
http://leap.ee.iisc.ac.in/sriram/teaching/MLSP$_1$6/$refs/GMM_Tutorial_Reynolds.pdfGaussianM$
$DouglasReynolds$