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. To verify the significance of the results, I use the T-test. Finally I figure out that the optimal algorithm in internal-term is not necessarily the one that give the optimal external fitment.

1 INTRODUCTION

The datasets I use are: *Online Shoppers, Diabetic-Data, and E-Shop Clothing 2008*. The goal of the research goal is to provide an optimal clusters set for each dataset. First, I transform the non-numeric values to numeric. To analyze the models in low-dimension I employ the *PCA* algorithm. For all the datasets I sample 11,000 data points. I use five different algorithms, to decide which of them is optimal I use a statistical test called T-test. I find which algorithm is the optimal both in external-term and internal-term.

2 METHODS

2.1 UPDATE 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.

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 I build a dictionary that holds these values with numeric keys 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

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 CLUSTERING ALGORITHMS

Clustering is an unsupervised machine learning algorithm family. It helps in grouping data points. In the following, I describe few of these algorithms.

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.

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.
end while

2.3.2 GMM- GAUSSIAN MIXTURE MODEL

Gaussian Mixture Model is a soft clustering algorithm. This is a soft-clustering approach since every point has a 0-1 chance to be picked for each cluster. The chosen cluster is the one with the highest probability. GMM is also an Expectation-Maximization algorithm.

The Algorithm (pseudocode):

The Algorithm:

Start with random Gaussian parameters (θ)
The center does not change anymore. : 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. It is applied to wide range of problems connected with feature analysis, clustering and classifier design. FCM clustering is an iterative process. The process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.

2.3.4 SPECTRAL CLUSTERING

Spectral clustering is an algorithm clustering based on graph theory. The main approach is 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.
We have more than one cluster :Take the two closest clusters and make them one cluster.

2.4 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.

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 initialization. The randomness can affect the results, to minimize the effect and 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 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.

2.7 OPTIMAL NUMBER OF CLUSTERS

All the cluster algorithms I use in my research have random initialization. I sample data points from the data. To provide the optimal number of clusters I run Silhouette 2.4 five times with different sample points each time. Finally, I use two-sample one-tailed T-Test 2.5 to decide which is the optimal.

2.8 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.

3 RESULTS

In the following, I describe the results for each dataset.

3.1 DATA-SET 1: ONLINE SHOPPERS INTENTION

The first data-set consists information about users in a one-year period. It contains 12,330 instances and 18 features.

3.1.1 SILHOUETTE VISUALIZATION

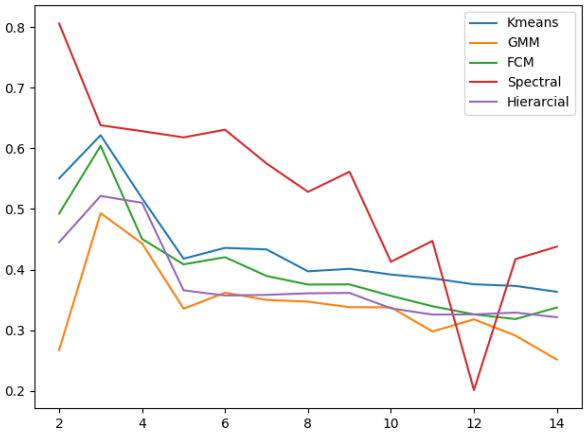

Figure 1: The maximum of Silhouette score for K-Means, FCM, Hierarchial and GMM received with 3 clusters and for Spectral 2 clusters. The method with the highest Silhouette score is Spectral.

3.1.2 CLUSTERING VISUALIZATION

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

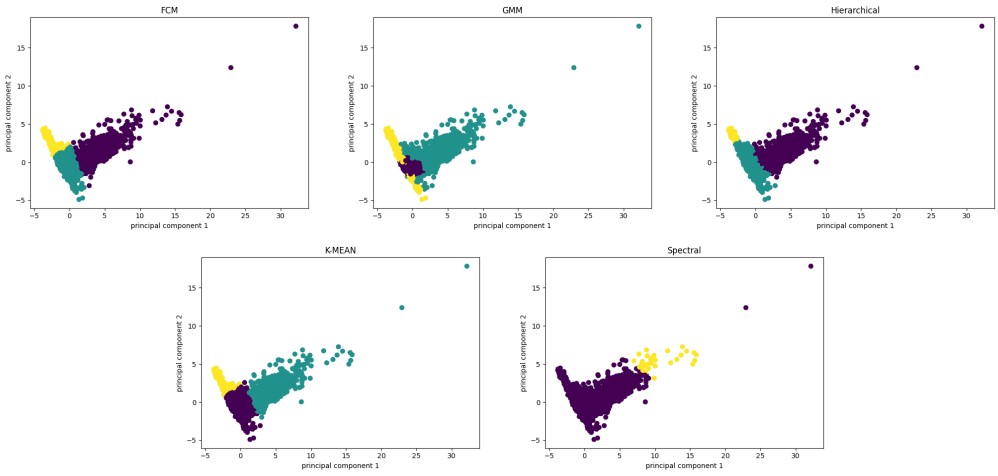

Figure 2: Data Set 1: Optimal clustering

3.2 DATA-SET 2: DIABETIC DATA

The second dataset contains 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It has 10,000 instances and 55 attributes.

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.

3.2.1 SILHOUETTE VISUALIZING

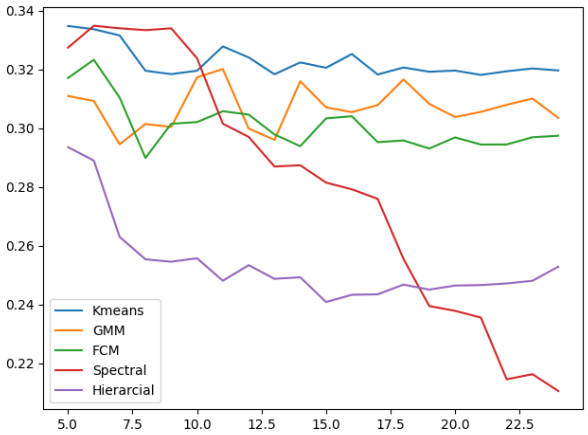

Figure 3: The maximum of Silhouette score for FCM and Spectral recived with 7 clustering, Hierarchial, GMM and K-Means received maximum with 5 clusters. The method with the highest Silhouette score is K-Means

3.2.2 CLUSTERING VISUALIZATION

In FCM, Spectral, and GMM are described in 2.3.2 the optimal number of clusters is corresponds to the Silhouette output. The method described in 2.7 determine that the optimal number of clustering for K-Means is 12 while it is not the output of the Silhouette.

3.2.3 EXTERNAL QUALITY OF CLUSTERS

3.3 DATA SET 3: E-SHOP CLOTHING 2008

The third data-set contains information on click stream 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.

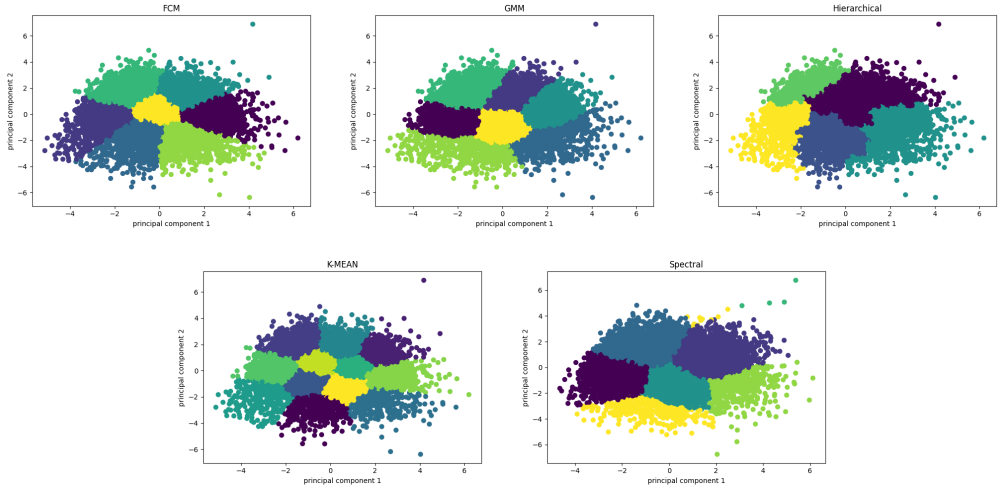

Figure 4: Data Set 2: Optimal clustering

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 produce a low value. The reason explained in section *4*. The cluster with the highest AMI is FCM, therefore it is the optimal cluster algorithm.

3.3.1 SILHOUETTE VISUALIZING

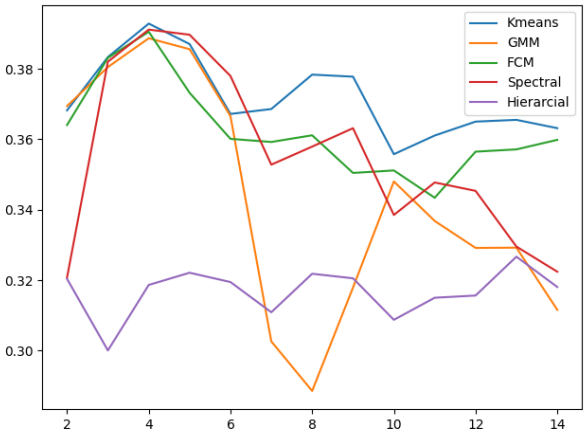

Figure 5: The maximum of Silhouette score for K-Means, GMM, FCM and Spectral received with 4 clusters but for Hierarchical Clustering received with 14. The cluster with highest Silhouette Score is K-Means.

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.

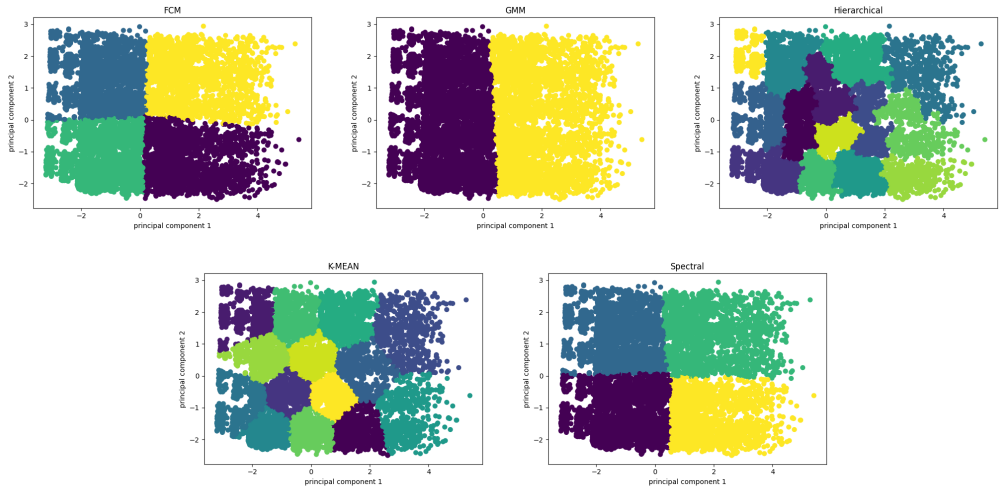

Figure 6: Data Set 3: Optimal clustering

3.3.3 EXTERNAL QUALITY OF CLUSTERS

Data 3- Adjusted Mutual 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: All the algorithms produce a low value. The reason explained in section 4. The cluster with the highest AMI is hierarchical. Therefore the optimal clustering algorithm to the given classification is Hierarchical.

4 DISCUSSION

In this work, I try to answer the question of which clustering algorithm is the best one for a specific task. First, as we seen this question is broad, I show 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 low in all the algorithms because the external classification is not correlated to the internal data. As we can see in the table below 2.3.1 gives the best clustering in internal-term for two out of three datasets, while 2.3.3 gives the best external fitment also for two out of three. We can conclude that 2.3.1 gives the best internal fitment in most of our cases and 2.3.3 produce 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$
$DouglasReynolds$