Comparison Between Five Clustering Algorithms

Ido Terner
January 2021

Abstract

Clustering 3 different data sets using **Clustering Algorithms**, showing results and analyzing performances for each, in an attempt to find the best clustering algorithm and anomaly detection method for a given data set. Finding the *K-Means algorithm* to be the best preforming algorithm in term of external and internal measures for the given data sets for the given data sets. All the code for this research can be find in GitHub
https://github.com/iTerner/UnsupervisedLearning

1 Introduction

Clustering is one of the most widely used techniques for exploratory data analysis. Its goal is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other and to cluster the data, you need to use a clustering algorithm. There are many cluster algorithms based on different assumptions on the data such as the number of clusters, the distribution of the data, shape, etc. In this research, we will present five different algorithms, **K-Means**, **Fuzzy-C-Means**, **Gaussian Mixture Model** also known as **GMM**, **Agglomerative Clustering** and **Spectral Clustering**, To check if there are any differences between those algorithms, and if so, try to prove, using statistical tests, that one algorithm is better than the others.

2 Methods

In this research, we use five clustering algorithms as well as different evaluation methods and statistical tests. In this section, we describe each of the methods we use.

2.1 Clustering Algorithms

In this section we describe the unique features of each clustering algorithm we use, and their approach to the clustering problem.

2.1.1 K-Means

K-Means is one of the most simplest algorithms in unsupervised learning, that look for predefined number of centeroids points in the data space that minimize the sum of the distances from every data point to her closest centeroid point. K-Means objective can find clusters that are convex and isotropic, which it is not always the case. The algorithm responds poorly to irregular shapes or manifolds clusters.

2.1.2 GMM

The Gaussian Mixture Model, also known as GMM, is a probabilistic model that attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input data set. GMM is a soft clustering algorithm, meaning it does not assign each point in the data to a cluster, but rather gives the probability for been in a specific cluster.

2.1.3 Fuzzy-C-Means

Fuzzy-C-Means is an algorithm that using the *Fuzzy Clustering*, which is a form of clustering in which each data point can belong to more than one cluster. The algorithm is similar to the K-Means algorithm, it selects a predefined number of clusters, and assign coefficients randomly to each data point for being in the clusters. The algorithm repeats itself until it converged, meaning the coefficients change between two iterations is no more than ϵ, the given sensitivity threshold.

2.1.4 Agglomerative Clustering

Agglomerative Clustering is one of two different types of *Hierarchical Clustering*. In this algorithm, data points are clustered using a bottom-up approach starting with individual data points. The algorithm form a cluster by joining the two closest points and form more cluster by joining the two closest clusters. The algorithm repeats those steps until one big cluster informed, and then, the *dendrograms* are used to divide into multiple clustering.

2.1.5 Spectral Clustering

Spectral Clustering is one of graph-based clustering, and in particular, an algorithm based on a *similarity graph*. The algorithm creates the similarity graph from the data, then computes the first k eigenvectors of its Laplacian matrix to define a feature vector for each object(where k is the predefined number of clusters), and finally, run k-means on these features to separate objects into k classes. The algorithm is considered to be expensive, due to the calculation of the eigenvalues and eigenvectors of the Laplacian.

2.2 Dimension Reduction

Working with high dimensional data can be hard, because of the computational cost of optimization in many dimensions and the *Curse Of Dimensionality* that make it difficult to gather insight without being 'tricked' by the many dimension. Therefore, it could be helpful to us reduce the dimensions of the data samples before doing the clustering.

2.2.1 PCA

Principal Component Analysis, also known as PCA, is a dimension reduction method that used to transform data into a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. PCA also helps us to visualize the data after the classification, by reducing the data to two-dimension.

2.3 Clustering Evaluation

To evaluate how well the algorithm cluster the data, we use several well known clustering evaluation methods.

2.3.1 Adjusted Mutual Information

Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is more information shared. The method returns a value between 0 to 1, where: 1 means that the two partitions are identical.

2.3.2 Silhouette Score

Silhouette Score is a method to evaluate the *internal clustering* by providing a succinct graphical representation of how well each object has been classified. The Silhouette Score is computed using the mean of the distances in each cluster (a) and the minimum distance from one cluster to another (b), the distance from a point to the nearest cluster that the

point is not a part of. The score for a point is: $\frac{b-a}{max(a,b)}$. The score can be $-1 < s < 1$ where 1 is the best value, $s \approx 0$ indicates overlapping clusters, and $s < 0$ indicates that the point is assigned to the wrong cluster.

2.4 Statistical Tests

The comparison between the different results can be statistically significant only if they are results not random. To test how likely it is to say that there is a difference between the results is not random we use statistical tests.

2.4.1 One Way ANOVA

Analysis of variance, also known as ANOVA, is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

2.4.2 Two-Sample One-Tailed T-Test

The Two-Sample One-Tailed T-Test is an alternative way of computing the statistical significance of a parameter inferred from a data set, in terms of a test statistic. The hypothesis H0 for A, b is 'the mean of A is greater than the mean of B'.The statistical test gives you a probability which marked by p. The *p value* is the probability of obtaining test results at least as extreme as the results observed, under the assumption that the null hypothesis is correct.

2.5 Data Sets

In this section, we will present the data sets we used in this research.

2.5.1 Online Shoppers Intention Data Set

The Online Shoppers Purchasing Intention data set contains information of 12,330 online shoppers encounters and whether they ended up with a sale. Of the 12,330 sessions in the data set, 10,422 were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping. The data set consists of feature vectors belonging to the different sessions

2.5.2 Diabetes Data Set

The data set represents 10 years of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. The Information contains only information related to diabetes. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, etc. Before using the data we explored it, cleaned it, and fix missing values.

2.5.3 Online Shopping Click Stream Data Set

The data set contains information on click-stream from an online store offering clothing for pregnant women. The 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.

3 Results

We cluster each data set with all the clustering algorithms as we mentioned before and For every algorithm, we try different hyper-parameters, report the results in addition to their

p-value. We uses the same number of clusters for all of the algorithm. We chose the number by looking at the average silhouette score of the and performing statistical tests.

Number Of Clusters	Shoppers Intention	Diabetes	Shopping Click
2	0.391	0.37	0.382
3	0.455	0.373	0.409
4	0.436	0.364	0.424
5	0.413	0.333	0.392
6	0.394	0.355	0.369
ANOVA p value	¡ 0.001	¡ 0.001	¡ 0.001

Table 1: The results of the ANOVA test for each data set.

As you can see in Table 1, The mean of each group is different, indicating that there is an optimal number of clusters. To find that number, we will perform a Two-Sample One-Tailed T-Test.

Group	Group	H0	p value
3	2	3 clusters are better then 2 clusters	≈0.99
3	4	3 clusters are better then 4 clusters	≈0.97
3	5	3 clusters are better then 5 clusters	≈0.99
3	6	3 clusters are better then 6 clusters	≈0.98

Table 2: The statistical tests and their p-value results for the Online Shoppers Intention Data Set.

Group	Group	H0	p value
3	2	3 clusters are better then 2 clusters	≈0.99
3	4	3 clusters are better then 4 clusters	≈0.97
3	5	3 clusters are better then 5 clusters	≈0.99
3	6	3 clusters are better then 6 clusters	≈0.98

Table 3: The statistical tests and their p-value results for the Diabetes Data Set.

Group	Group	H0	p value
3	2	3 clusters are better then 2 clusters	≈0.99
4	3	4 clusters are better then 3 clusters	≈0.97
4	5	4 clusters are better then 5 clusters	≈0.99
4	6	4 clusters are better then 6 clusters	≈0.98

Table 4: The statistical tests and their p-value results for the Online Shopping Click Stream Data Set.

By looking at Table 2 we can see that the optimal number of clusters for the Online Shoppers Intention Data Set is 3. According to Table 3, the optimal number of clusters for the Diabetes Data Set is 3. Moreover, by looking at Table 4, we can see that the optimal number of clustering for the Online Shopping Click Stream Data Set is 4.

We ran each algorithm 30 times on each data set. For the Online Shoppers Intention Data Set, we use 3 clusters, for the Diabetes Data Set we use 3 clusters, and for the Online Shopping Click Stream Data Set, we use 4 clusters. Then, we compute the Average silhouette score for each algorithm as you can see in Tables 5, 6 and 7.

In order to check how well the data correlate with the clusters, we will compute the average Adjusted Mutual Information of each data set with the given tag. In the Online Shoppers Intention Data Set, we have 3 tags, therefore, we will compute the Adjusted Mutual Information 3 different times. we can see in Table 8 the average Adjusted Mutual Information for each tag. We can see that the data is not correlated with the tags In the Diabetes Data

Algorithm	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Average Score	0.455	0.4496	0.4544	0.4135	0.454

Table 5: The Average silhouette score of the Online Shoppers Intention Data Set.

Algorithm	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Average Score	0.3733	0.3716	0.3721	0.3175	0.3735

Table 6: The Average silhouette score of the Diabetes Data Set.

Algorithm	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Average Score	0.42	0.417	0.418	0.37	0.417

Table 7: The Average silhouette score of the Online Shopping Click Stream Data Set.

Set, we have 2 tags, therefore, we will compute the Adjusted Mutual Information twice. we can see in Table 9 the average Adjusted Mutual Information for each tag. We can see that the data is not correlated with the tags. In the Online Shopping Click Stream Data Set, we have one tag therefore, we will compute the Adjusted Mutual Information only once. We can see in Table 10 the average Adjusted Mutual Information of the tag, we can infer that the data is not correlated with the tag.

Classifier	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Revenue	0.03105	0.03347	0.031	0.031	0.0305
Visitor Type	0.01096	0.0094	0.0103	0.01	0.01
Weekend	0.151	0.1332	0.147	0.1024	0.142

Table 8: The Average Adjusted Mutual Information of the Online Shoppers Intention Data Set tags.

Classifier	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Gender	0.0017	0.0017	0.0017	0.0017	0.0001
race	5.424e-5	0.0001	5.717e-5	1.549e-5	3.33e-5

Table 9: The Average Adjusted Mutual Information of the Diabetes Data Set tags.

Classifier	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Gender	0.029	0.0259	0.0277	0.0002	2.886e-5

Table 10: The Average Adjusted Mutual Information of the Online Shopping Click Stream Data Set tag.

To determine which cluster algorithm is the best for each data set, we will perform the Two-Sample One-Tailed T-Test. As you can see in Table 11, the best clustering algorithm for the Online Shoppers Intention Data Set is K-Means. According to Table 12, we can see that the best algorithm for the Diabetes Data Set algorithm is K-Means. And, as we can see in Table 13, the algorithm with the best performance is K-Means, but, except for the Agglomerative Clustering, all the algorithms had a similar average. The reason that K-Means had the best performance among the other algorithms might be due to the high density at which the data points are compressed, which makes most of the data points that are in the same cluster very close to each other.

Algorithm	Algorithm	H0	p value
K-Means	GMM	K-Means is better than GMM	≈0.99
K-Means	Fuzzy-C-Means	K-Means is better than Fuzzy-C-Means	≈0.99
K-Means	Agglomerative	K-Means is better than Agglomerative	≈1
K-Means	Spectral	K-Means is better than Spectral	≈0.98

Table 11: The statistical tests and their p-value results for the Online Shoppers Intention Data Set.

Algorithm	Algorithm	H0	p value
K-Means	GMM	K-Means is better than GMM	≈0.98
K-Means	Fuzzy-C-Means	K-Means is better than Fuzzy-C-Means	≈0.92
K-Means	Agglomerative	K-Means is better than Agglomerative	≈0.98
K-Means	Spectral	K-Means is better than Spectral	≈0.61

Table 12: The statistical tests and their p-value results for the Diabetes Data Set.

Algorithm	Algorithm	H0	p value
K-Means	GMM	K-Means is better than GMM	≈0.72
K-Means	Fuzzy-C-Means	K-Means is better than Fuzzy-C-Means	≈0.58
K-Means	Agglomerative	K-Means is better than Agglomerative	≈0.99
K-Means	Spectral	K-Means is better than Spectral	≈0.6

Table 13: The statistical tests and their p-value results for the Online Shopping Click Stream Data Set.

4 Summary

In this research, we compared different clustering algorithms with different data sets. Even though the data sets were very different from each other, we recognize some patterns. The first is that K-Means performed better than all the other clustering algorithms, and so the Spectral Clustering. The second is that the data was not correlated with the external tags, indicating that they might capture different patterns in the data.

References

[1] Trupti M. Kodinariya, Dr. Prashant R. Makwana. Review on determining number of Cluster in K-Means Clustering. 2013

[2] Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, 2016

[3] Michal Valko, Graphs in Machine Learning,
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.708.6592rep=rep1type=pdf