COMPARATIVE ANALYSIS OF UNSUPERVISED LEARNING METHODS

Ido Terner
March 2021

Abstract

Clustering 2 different data sets using **Clustering Algorithms**, comparing between different **Anomaly Detection Methods**, 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, and *Cluster-Based Outlier Detection* as the best anomaly detection method for the given data sets. All the code for this research can be find in GitHub

https://github.com/iTerner/UnsupervisedLearning

1 Introduction

Data clustering is a procedure in which we make cluster of entities ton based on their similar features. A good clustering technique will create high quality clusters with high intra-class similarity low inter-class similarity Quality of clustering depends on the similarity measure used and its implementation. 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. Moreover, we will use *Anomaly Detection* techniques such as **Cluster-Based outlier detection** and **Density-Based outlier detection** to find the anomalous points, try to determine which method has the best impact on the data sets.

2 Methods

In this research, we use five clustering algorithms as well as different evaluation methods, statistical tests and anomaly detection methods. 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. In this research, we used the clustering algorithms to cluster the data and determine which algorithm is the best.

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 a 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 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.

2.4.2 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.

2.5 Anomaly Detection

In this section, we will describe the methods we use in order to detect the anomalous points from the data sets.

2.5.1 Cluster-based outlier detection

The Cluster-based outlier detection method uses the classification results and score of each data point. The method cluster the data using an algorithm (such as K-Means, GMM, etc) and compute the score of each data point. Then, we remove the data points with negative scores and repeat until coverage.

2.5.2 Density-based outlier detection

The Density-based outlier detection method investigates the density of an object and that of its neighbors. The idea of density-based is that we need to compare the density around an object with the density around its local neighbors. The basic assumption of density-based outlier detection methods is that the density around a non-outlier object is similar to the density around its neighbors, while the density around an outlier object is significantly different from the density around its neighbors.

2.6 Data Sets

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

2.6.1 Mo-Cap Hand Postures Data Set

Mo-Cap Hand Postures Data Set contains instances which are unorderd cloud of 3D points representing one of five hand postures. The data was gathered with few candidates that held their hand in one of the target postures wearing special glove that with markers at different points along the glove. Special camera took the picture of the posture and with simple computation the markers normalized to 3D point representing relative position to the other points. To fill the empty features, we filled the missing data with the median of the column.

2.6.2 HTRU2 Data Set

HTRU2 is a data set which describes a sample of pulsar candidates collected during the high Time Resolution Universe Survey. Pulsars are a rare type of neutron star that produce radio emission detectable here on earth. They are of considerable scientific interest as probes of space-time, the interstellar medium, and states of matter. The data set contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators.

3 Results

3.1 Clustering

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 and detect the anomalous points in the data set. 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	Mo-Cap	HTRU2
2	0.388	0.401
3	0.434	0.439
4	0.406	0.445
5	0.399	0.439
6	0.4	0.424
ANOVA p value	< 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	$\approx$0.99
3	4	3 clusters are better then 4 clusters	$\approx$0.97
3	5	3 clusters are better then 5 clusters	$\approx$0.99
3	6	3 clusters are better then 6 clusters	$\approx$0.98

Table 2: The statistical tests and their p-value results for the Mo-Cap Hand Postures Data Set.

By looking at Table 2 we can see that the optimal number of clusters for the Mo-Cap Hand Postures Data Set is 3, and according to Table 2 the optimal number of clusters for the HTRU2 Data Set is 4.

We ran each algorithm 30 times on each data set. For the Mo-Cap Hand Postures Data Set we use 3 clusters and for the HTRU2 data set we use 4 clusters. Then, we compute the Average silhouette score for each algorithm as you can see in Table 4 and 5.

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	≈0.99
4	5	4 clusters are better then 5 clusters	≈0.99
4	6	4 clusters are better then 6 clusters	≈0.98

Table 3: The statistical tests and their p-value results for the HTRU2 Data Set.

Algorithm	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Average Score	0.445	0.380	0.444	0.375	0.432

Table 4: The Average silhouette score of the Mo-Cap Hand Postures Data Set.

Algorithm	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
Average Score	0.434	0.381	0.43	0.393	0.428

Table 5: The Average silhouette score of the HTRU2 Data Set.

In order to check how well the data is correlated with the clusters, we will compute the average Adjusted Mutual Information of each data set with the given tag. In the Mo-Cap Hand Postures Data Set, we have one tag therefore, we will compute the Adjusted Mutual Information only once. We can see in Table 6 the average Adjusted Mutual Information of the tag. By looking at the average score, we can infer that the external quality of clustering reasonable, meaning that the correlation of the data is not the best but neither the worst. In the HTRU2 Data Set, we also have one tag and therefore, we will compute the Adjusted Mutual Information only once. As we can see in Table 7 .we can see that our external quality is not good, indicating that the clusters were not correlated with the tag.

Classifier	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
class	0.43	0.455	0.431	0.402	0.424

Table 6: The Average Adjusted Mutual Information of the Mo-Cap Hand Postures Data Set tag.

Classifier	K-Means	GMM	Fuzzy-C-Means	Agglomerative	Spectral
class	0.146	0.148	0.141	0.037	0.163

Table 7: The Average Adjusted Mutual Information of the HTRU2 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 8, the best clustering algorithm for the Mo-Cap Hand Postures Data Set is K-Means. And, by looking at Table 9, we can see that the best algorithm for the HTRU2 algorithm is K-Means. The reason for that is 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, and therefore, K-Means results give the highest average score among the algorithms.

3.2 Density Estimation and Anomaly Detection

As we describe earlier, we want to determine which anomaly detection method is better, the Cluster-based outlier detection or the Density-based outlier detection. To do that, we will remove all the anomalous points for each data set and compute the average silhouette score of each algorithm. As we can see in Table 10 and Table 11, There is a significant difference between the average silhouette score of each algorithm W/O any anomaly detection method.

Algorithm C height	Algorithm	H0	p value
K-Means	GMM	K-Means is better than GMM	≈1
K-Means	Fuzzy-C-Means	K-Means is better than Fuzzy-C-Means	≈0.97
K-Means	Agglomerative	K-Means is better than Agglomerative	≈1
K-Means	Spectral	K-Means is better than Spectral	≈0.98

Table 8: The statistical tests and their p-value results for the Mo-Cap Hand Postures Data Set.

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

Table 9: The statistical tests and their p-value results for the HTRU2 Data Set.

Also, after performing some statistical tests to determine which method is better. By looking at Table 12, we can see that the Cluster-Based outlier detection is much better than the Density-Based outlier detection. The reason for that might be the fact that the data points are very crowded, therefore, most of the points have high-density value, thus, the Density-Based outlier detection performed very poorly for those data sets.
,

Algorithm	Without any detection method	Cluster-Based	Density-Based
K-Means	0.445	0.531	0.471
GMM	0.38	0.531	0.426
Fuzzy-C-Means	0.444	0.528	0.465
Spectral	0.432	0.518	0.458
Agglomerative	0.375	0.525	0.432

Table 10: The average silhouette score of the Mo-Cap Hand Postures Data Set W/O an anomaly detection methods.

4 Summery

In this research, we compared different clustering algorithms and anomaly detection methods with variant data sets. Although the data sets were different from each other, we recognize some patterns. The first is that K-Means had the best performance among all the other algorithms. Second, the Cluster-Based outlier detection performed better than the Density-Based outlier detection, indicating that the data might be crowded. Moreover, most of the data was not correlated with the external tags, indicating that they might capture different patterns in the data.

Algorithm	Without detection method	Cluster-Based	Density-Based
K-Means	0.434	0.564	0.457
GMM	0.38	0.554	0.393
Fuzzy-C-Means	0.43	0.557	0.456
Spectral	0.428	0.552	0.44
Agglomerative	0.393	0.558	0.41

Table 11: The average silhouette score of the HTRU2 Data Set W/O an anomaly detection methods.

Method	Method	H0	p value
Cluster-Based	Density-Based	Cluster-Based is better than Density-Based	≈0.99
Cluster-Based	Density-Based	Cluster-Based is better than Density-Based	≈0.99

Table 12: The statistical tests and their p-value results for the Mo-Cap Hand Postures(1^{st} row) and the HTRU2(2^{nd} row) Data Sets.

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