Introduction to Machine Learning

Aleix Solanes, Pablo Martinez
Master in Artificial Intelligence

Abstract

Clustering is the process of organizing objects into different groups whose elements are similar in some way. The goal of clustering is to determine the grouping in a set of unlabeled data. This first work of Introduction to Machine Learning, has the aim of analyzing different clustering algorithms by using a few different data sets. The algorithms that will be analyzed are K-Means, Bisecting k-means and Fuzzy C-Means.

I. Methods

The three algorithms that have been implemented are K-means, Bisecting K-means and Fuzzy C-Means. In the following section this algorithms will be detailed.

I. Clustering algorithms

I.1 K-Means

K-means is a simple unsupervised learning algorithm that assuming a number of K clusters fixed a priori, classifies a given data set through the k specified clusters. The algorithm is as follows:

1. Randomly place K points into the space represented by the objects that are being clustered. These points can be considered as the initial centroids of each group.
2. Assign each object to the group that has the closest centroid.
3. Once all the objects have been assigned to a cluster, recalculate the position of each centroid.
4. Repeat the steps 2 and 3 until the centroids no longer move.

I.2 Bisecting K-Means

Bisecting K-means is a different approach to clustering a data set. It can be considered as a derivation of the K-means algorithm. The algorithm is as follows:

1. The algorithm considers an initial single global cluster.
2. Decide which cluster to split (for example by choosing the biggest cluster available), and split it by using the K-means algorithm.
3. Repeat the second step for n times, and take the split that produces the clustering with the highest overall similarity.
4. Repeat the steps 1, 2 and 3 until the number of clusters available are equal to the a priori specified K.

I.3 Fuzzy C-means

Fuzzy C-means is a method of clustering that allows one instance of data to belong to more that one cluster. Thus, points in the edge of a cluster can belong to the cluster in a lesser degree as the ones in the center. The algorithm is as folows:

1. Randomly select cluster center.
2. Initialize U matrix
3. At k-step: calculate the centers of the vectors
4. Update the U matrix

5. Repeat 3 and 4 until it converges. The convergence will occur when the coefficients' change between two iterations are less than epsilon (the given sensivity threshold).

II. Validation metrics

In order to validate the results obtained from the clustering algorithms, three indexes of quality will be used. These metrics, are evaluated against external data, in this case known class labels (ground truth). The methods used are explained below.

II.1 Purity

The purity index, is a simple measure, that consists in the ratio between the dominant class in the cluster, and the size of the cluster.

II.2 Rand Index

It is a measure of similarity between two data clusterings. The simplified formula consists in the following variables:

- **a**: the number of pairs of elements that are in the same class both in P and G.
- **b**: the number of pairs of elements that are in the same class in P but different in G.
- **c**: the number of pairs of elements that are in different classes in P but same in G.
- **d**: the number of pairs of elements that are in different classes both in P and G.

The formula is as follows:

$$\frac{a+d}{a+b+c+d} \tag{1}$$

II.3 Adjusted Rand Index

The adjusted Rand Index is the normalized difference of the Rand Index and its expected. The formula is as follows:

$$\frac{RandIndex - ExpectedRandIndex}{1 - ExpectedRandIndex} \tag{2}$$

II.4 F-measure

The F-measure or F-score, is a measure of the accuracy of a test. It considers two values, the precision and the recall, in order to compute a score. Considering that the parameters are:

- **Precision (P)**

$$\frac{TruePositives}{TruePositives + FalsePositives} \tag{3}$$

- **Recall (R)**

$$\frac{TruePositives}{TruePositives + FalseNegatives} \tag{4}$$

Then, the formula is as follows:

$$\frac{2 * P * R}{P + R} \tag{5}$$

II. Results

The data to be analyzed will be three data sets. The previously descrived algorithms have been applied to each dataset, and the validation metrics explained will be used in order to verify the quality of each result.

I. Data sets

The datasects selected to be evaluated are the following:

- **ionosphere**
- **wine**
- **iris**

II. 1st data set: ionosphere.arff

This data set contains 2 classes and 351 instances. This system consists of a phased array of 16 high-frequency antennas. The targets are free electrons in the ionosphere. The classes are: "Good", which refers to electrons that show the evidence of some type of structure in the ionosphere, and "Bad", those that did not show any evidence of structure.

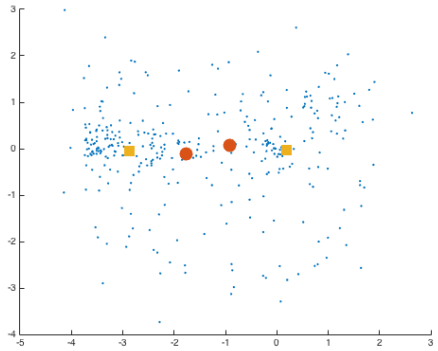

Figure 2: *K-means optimal clustering*

Table 1: *Confusion matrix of the KM*

	Good	Bad
Cluster 1	157	68
Cluster 2	33	93

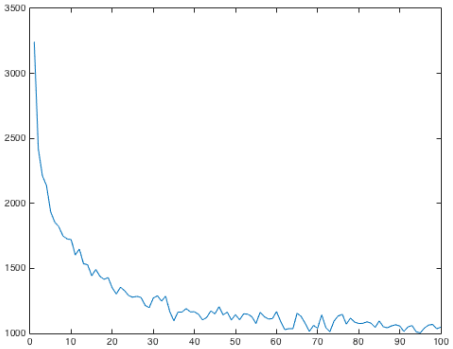

Figure 3: *K-means optimal clustering*

Table 2: *Confusion matrix of the FCM*

	Good	Bad
Cluster 1	92	34
Cluster 2	68	157

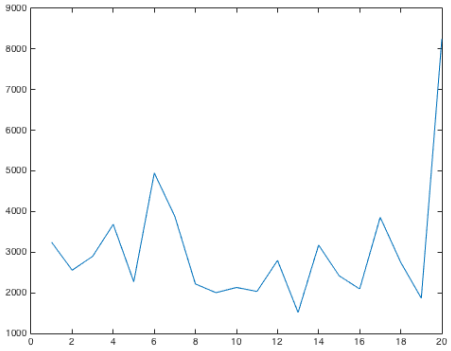

Figure 1: *K-means optimal clustering*

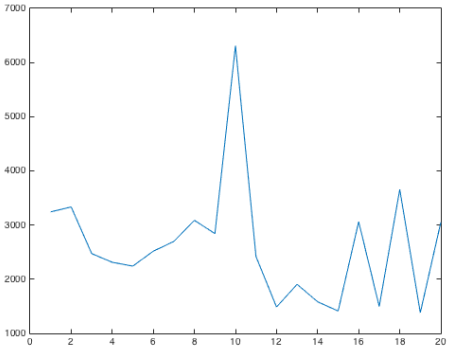

Figure 4: *Fuzzy C-means optimal clustering*

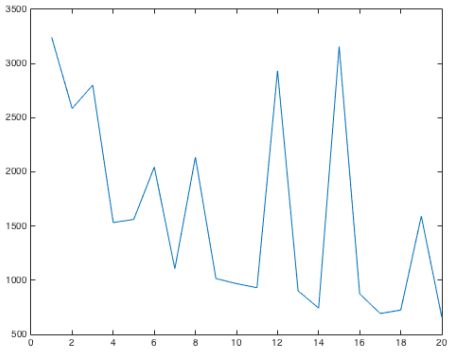

Figure 5: *Fuzzy C-means optimal clustering*

Considering that the number of clusters predefined are the same as the ground truth has, the validation metrics give the following table.

Table 3: *Evaluation of methods for K=2*

Method	KM	BKM	FCM
ARI	0.1776	0.1776	0.1727
Purity	0.7123	0.7123	0.7094

III. 2nd data set: wine.arff

This data is information of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. Each class corresponds to a different cultivar. The class 1 has 59 instances, class 2 71 and class 3 48.

In order to see the results of the Fuzzy C-means algorithm in the wine dataset, the next table represents the confusion matrix.

Table 4: *Confusion matrix of the FCM*

	cultivar 1	cultivar 2	cultivar 3
Cluster 1	27	21	0
Cluster 2	20	50	1
Cluster 3	14	0	45

In the following image, the yellow squares, correspond to the calculated centroids, and the orange circles, the original dataset. The original centroids are calculated as the average of all points in the original cluster.

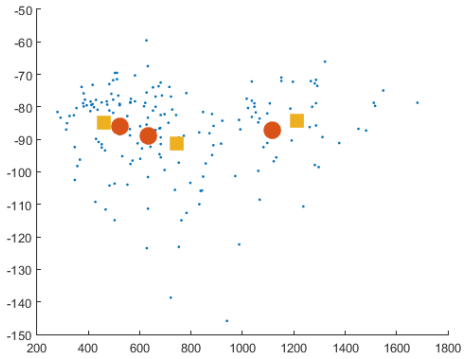

Figure 6: *Original clusters against calculated for wine.arff.*

By using the algorithm to determine the optimal number of clusters, the following images show the results depending on the number of clusters used.

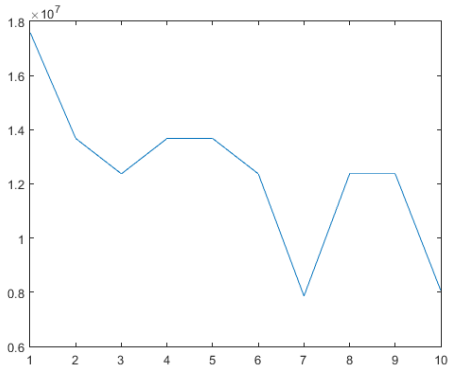

Figure 7: *K-means optimal clustering for wine.arff*

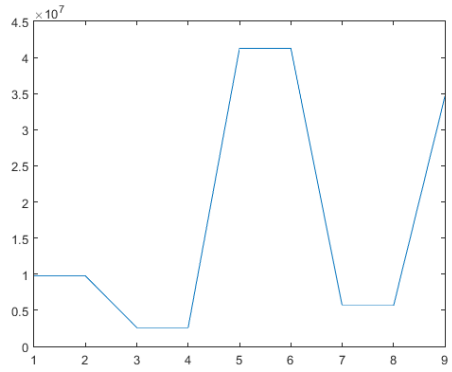

Figure 8: *Bisecting k-means optimal clustering for wine.arff*

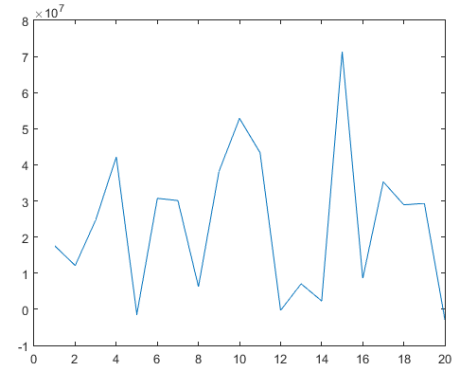

Figure 9: *Fuzzy C-means optimal clustering for wine.arff*

IV. 3rd data set: iris.arff

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2. The latter are not linearly separable from each other.

Table 5: *Confusion matrix of the FCM*

	setosa	versicolor	virginica
Cluster 1	45	0	5
Cluster 2	0	50	0
Cluster 3	8	0	42

Table 6: *Confusion matrix of the KM*

	setosa	versicolor	virginica
Cluster 1	45	3	2
Cluster 2	0	50	0
Cluster 3	14	0	36

Table 7: *Evaluation of methods for K=3*

Method	KM	BKM	FCM
ARI	0.7163	0.7209	0.7709
F-measure	0.8767	0.8933	0.9133

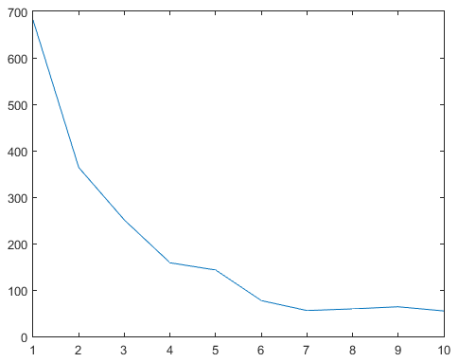

Figure 10: *Optimal clusters using KM for the iris.arff data set.*

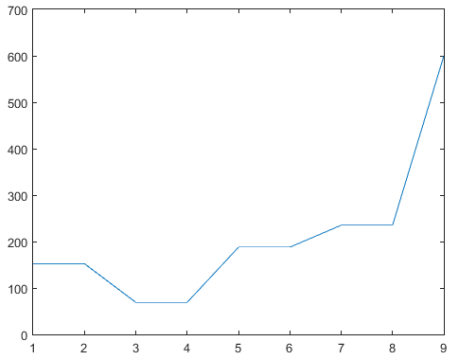

Figure 11: *Optimal clusters using BKM for the iris.arff data set.*

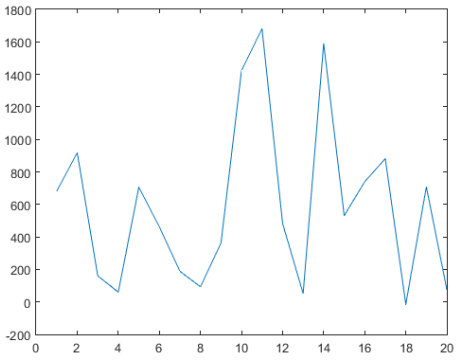

Figure 12: *Optimal clusters using FCM for the iris.arff data set.*

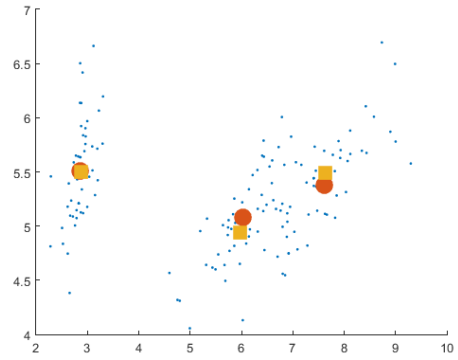

Figure 13: *Original clusters against calculated using FCM for the iris.arff data set.*

III. Discussion

I. Algorithms performance

In the datasets chosen in this work is hard to say which algorithm provides us the best overview of them, even that, Fuzzy C Means creates a really usfeul tool to evaluate a dataset, that tool is the partition matrix. It gives us a more exhaustive vision of how is the data organized.

II. Optimal K Value

In both K-Means and Bisecting K-Means is necessary to find an optimal K to work with. This problem has no analytical solution and a bunch of approximations can be found on the literature.

The implementation on this work is the brute force solution, it consists in calculate the square differences of each vector with the centroid of the class where it belongs to.

In the evaluated datasets the result of this calculus looks (In most of them) as an inverse exponential. In order to find the optimal K we've used the elbow rule, that consists in picking the value n that maximizes $f(n-1) - f(n)$ and minimizes $f(n) - f(n+1)$.

This gives a number that, in the case of Iris, which is a strongly clustered dataset, the number of real clusters. But, although, in the other datasets it gives a number that is far from being the real classes number.

References

[Bishop, Christopher] Pattern recognition and Machine Learning. *Springer*

[G.W. Milligan, M.C. Cooper] "An examination of procedures for determining the number of clusters in a data set." *Psychometrika, vol 50, 1985, pp. 159-179*

[Ka Yee Yeung, Walter L. Ruzzo] Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper âĂIJAn empirical study on Principal Component Analysis for clustering gene expression dataâĂİ (to appear in Bioinformatics) *May 3, 2001*