International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-6, Issue-1, September 2016

Determination of the number of cluster a priori using a K-means algorithm

Agustín Sáenz López, Facundo Cortés Martínez

***Abstract*— The K-means algorithm is currently one of the most important methods to obtain the cluster of a set of data, this method has as main problem, knowing the number of cluster that exist in the data set, before carrying out the analysis on the data set, however this parameter is usually not known. There are several methods to know a priori the number of cluster in the data set, none of them is decisive, but help to have an idea of the number of cluster for which you are searching. This article studies the application the elbow method to know the number of cluster in a set of data related to mixtures of concrete. It was found that this method mark between 5 and 9 cluster associated with this set of data.**

***Index Terms:*—Clustering, K-means Algorithm, Elbow Method, Number of Cluster.**

I. INTRODUCTION

The k-means clustering algorithm [1] is the grouping method more popular [2], is a method of unsupervised learning and its popularity is due to its simplicity with which it works, this algorithm has been used successfully in many branches of science. However the main problem that presents, is the need to determine in advance the number of cluster that exist in the set of data that you want to analyze. In the literature there are several jobs to determine the number of clusters a priori [3], however we will focus only on the method of the elbow.
The k-means algorithm is has the following process: For a data set $D:\{d_1, d_2, \cdots, d_n\}$ containing n objects and a value of k given:
1.- Select arbitrarily k data from the data set D as initial centroids.
2.- to associate each of the data d_i to the nearest centroid considering the Euclidean distance.
3.- Calculate the new average value of the d_i associated with each centroid and therefore to each cluster.
4.- If there is change in the centroids repeat step 2, with the new centroids.

The K-mean clustering algorithm takes the input parameter k, and partitions a set of n objects into k clusters so that the similarities in the interior of the cluster is high, but that the similarity between the cluster is low. The similarity within the cluster is measured considering the Euclidean Distance of the instances to the centroid of the cluster. Generally, the criterion of the square error is used as a measure of similarity and is defined as

$$E = \sum_{i=1}^{k} \sum_{j \in C_i} |x_j - C_i|^2$$

Where E is the sum of the square error for all objects in the data set; x_j is the point in space representing a given object; and C_i is the average of the cluster i (both x_j and C_i are multi-dimensional). In other words, for each object in each cluster, the distance of the object to its cluster is square and the distances are summed.

II. BACKGROUND

Tibshirani et al [4] developed a method to determine the number of clusters on the basis of the statistics of gap, the technique uses clusters obtained by the k-means clustering algorithm and compares the change with the dispersion that exists within the cluster with respect to the expected dispersion with a distribution of reference data.

Koteswara and Sridhar Reddy [5] propose a heuristic method on the basis of the ordering and partition the data, to find the initial centroids in accordance with the distribution of the data. The method proposed to sort the input data set and partition the data already sorted in k clusters, the average values of each of these sets partitioned correspond to the initial centroids.

ISHIOKA [6] proposed a method that initially divides the data set into clusters whose number is small enough, and continuing the division of each clusters in two clusters. The criterion used for the division is the Bayesian Information, some features of this method are; this applies to the set of data in general or with p-dimension; it is considered the variance and covariance around the centers of the clusters, evaluates the number of clusters by means of computer simulation.

Pelleg and Moore [7] propose an algorithm to find the optimal number of clusters, this number optimum clusters are should be between a minimum number of clusters which is the lower bound and a number of clusters maximum that is the higher bound, begins with the minimum number of centroids and continuously adding centroids in where they are needed to reach the upper bound. The algorithm consists basically in two operations, the first runs the k-means algorithm with a given number of clusters until you reach the convergence and the other process decides where it has placed a new centroid on the basis of the breakdown of the clusters in two clusters.

Greg Hamerly and Charles Elkana [8] propose the G-means algorithm that begins with a single centroid and is creating new centroids if the points that form a cluster continue or not a

Agustin Saenz Lopez, Faculty of Engineering, Science and Architecture Juárez University of Durango State, Mexico. (FICA-UJED)
Facundo Cortes Martinez, Faculty of Engineering, Science and Architecture Juárez University of Durango State, Mexico. (FICA-UJED)

distribution of Gauss, the cluster or centroid is not changed but if it do not follow a Gaussian distribution then this cluster is broken into two centroids and then runs the k-means algorithm again.

III. ELBOW METHOD

The method of the elbow [9] consists in carrying out a broad sweep of the number of cluster using the algorithm k-mean, starting with k=2 and go increased this value up to a large number that we think has exceeded the number of cluster that we believe exists in the data set, in each one of the runs for each K we obtain the square error of the cluster with respect to the instances that are contained in each one of the cluster, this value is plotted with respect to the number of cluster and we can get a graph as shown in the following graph 1.

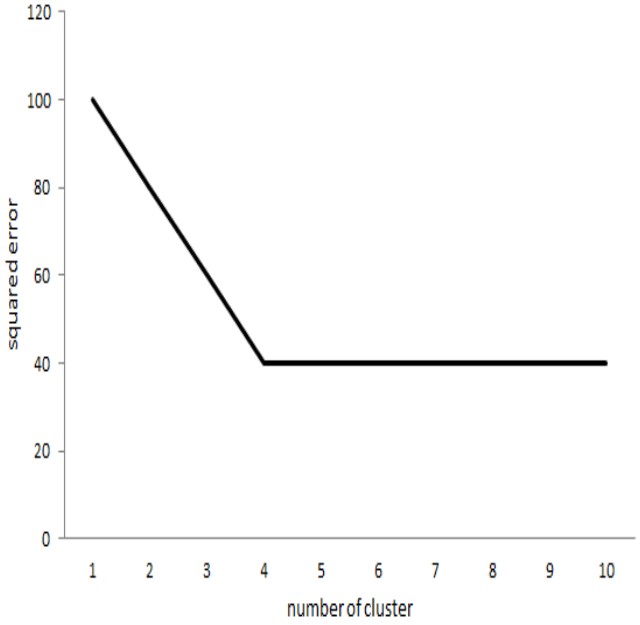

Fig. 1: Elbow point

In graph 1, we see that the square error decreases quickly to a point where this error remains more or less constant, The inflection point where it carries out this change, is the point elbow and is the number of cluster that this method tells us exist in the data set. However there are cases in which the graph that is obtained is as shown in Figure 2, where the decline of the square error is constant, and therefore we cannot carry out any determination on the number of existing cluster in the data set.

IV. METODOLOGY

For the experimental tests took the database of concrete mixtures that are found in the page of ICU, which consists of 1030 instances, where each instance has 9 attributes, which are cement, slag, ash, water, SP, coarse aggregate, fine aggregate, age, strength. All data are numerical

For the analysis of the data set with the K-mean algorithm, we use the software weka, there were a total of 14 tests, starting with a k value equal to 2 and ending with a k equal to 14 with increase of 1 cluster in each test.

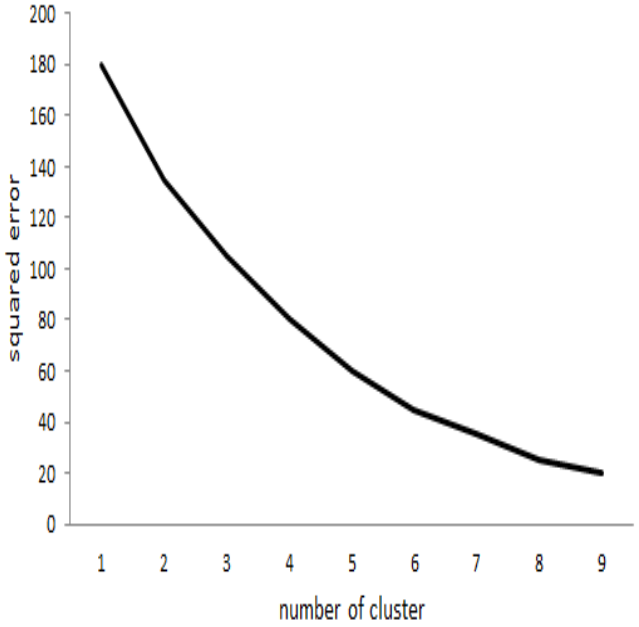

Fig. 2: ambiguity to identifying elbow point

V. RESULT AND DISCUSSION

To obtain the number of cluster that can contain the data set, obtained the graph of the square error of the software Weka against the set of cluster, which is shown in graph 3.

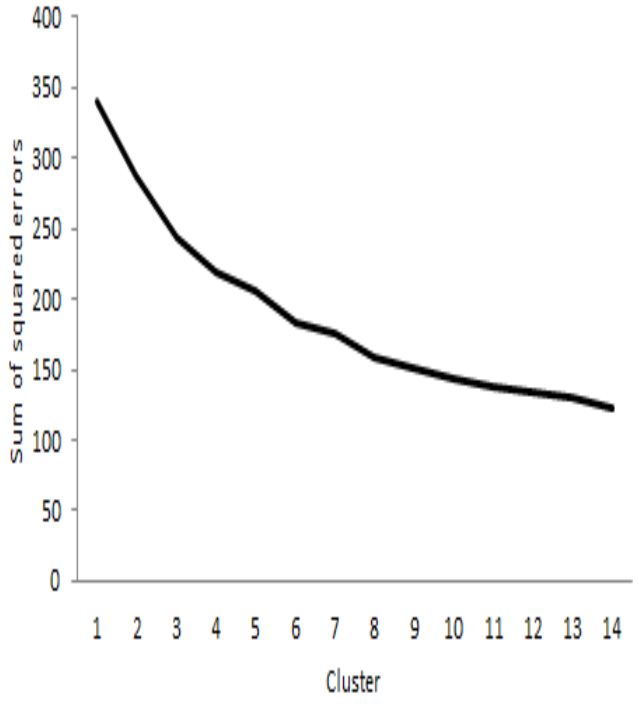

Fig 3: Sum of the squares errors against the number of clusters.

Due to that in the previous graph is not clear the number of cluster, we proceeded to draw the difference of the square error, obtaining the following graph.

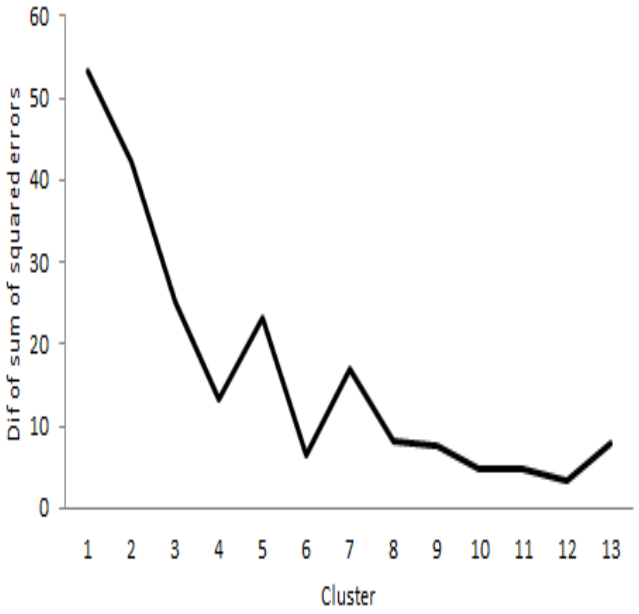

Fig 5: Difference of the sum of the squares errors against the number of clusters.

In the Graph 5 shows that there is a well defined descrease that starts from the cluster 2 until the cluster equal to 5 for after entering a zone of instability and then return to an area of continuous decline after the cluster 9, this second area of decrease of the difference of the error is constant but to a lesser extent than the first zone.

VI. CONCLUSION

In this work we applied the method of the elbow to a set of data to obtain a priori the number of cluster that might have this data set. The method of the elbow uses the graph of the distance of each object to the center of the cluster to which it is associated, however it was found that this graph we do not give sufficient information to get the number of cluster, so we use the difference of the squared error between two consecutive cluster and could already be observed a plot a bit better for our studio, where we conclude that this data set must be between 5 and 9 cluster associated with this set of data.

ACKNOWLEDGMENT

The authors are thankful to PIFI2014 for the support received for the realization this work

REFERENCES

[1] MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5th Berkeley Symp. Math. Statistics and Probability, 1:281-297, 1967

[2] R. C. Dubes and A. K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.

[3] Kodinariya and Makwan, Review on determining number of Cluster in K-Means Clustering, International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 6, November 2013.

[4] Robert Tibshirani, Guenther Walther and Trevor Hastie, Estimating the number of clusters in a data set via the gap statistic, J.R. Statist. Soc (2001),63, Part 2,pp 411-423.

[5] N. Koteswara Rao, G. Shridhar Reddy, Discovery of Preliminary Centroids using Improved K-Means Clustering Algorithm, International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4558-1561

[6] Tsunenori Ishioka, Extended K-means with an Efficient Estimation of the Number of Clusters, Ouyou toukeigaku Vol. 29 (2000) No 3 P 141-149.

[7] Dan Pelleg and Andrew Moore, X-means: Extending k-means with Efficient Estimation of the Number of Clusters, Proceeding of the Seventeenth International Conference on Machine Learning, pages 727-734.

[8] Greag Hamerly and Charles Elkan, Learning the k in k-means, Advances in neural information processing systems 17, March 2004.

[9] Andrew Ng. Clustering with K means Algorithm, Machine Learning, 2012