International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

# Using Clustering Algorithm to Determine the Number of Clusters

**Aditya Darak[1], Akhil Chaudhary[2], Prajwal Mogaveera[3]**

B.E (C.S.E.) Mumbai University, Mumbai, Maharashtra, India

**Abstract:** ***Clustering is important technique in data mining. The process of clustering involves partitioning of data into groups on the basis of similarities and differences between them. Clustering is used in various fields such as psychology, biology, data mining, image analysis, economics, pattern recognition, bioinformatics, weather forecasting, etc. The result of clustering varies as the number of cluster parameter changes. Therefore, the main challenge to cluster analysis is that the number of clusters or the number of parameters is seldom known and must be determined before clustering. Several clustering algorithms have been proposed. Among them the k-means clustering is a simple and fast clustering technique. Here, we address the problem of selecting the number of clusters by using a k-means approach. We can ask the end users to provide number of clusters in advance. But it may not always be feasible as the end user requires the domain knowledge of each data set. The initial cluster centres varies directly as the number of clusters. Thus, it is quite important for k-means to have good initial clusters. There are many methods available to estimate the number of clusters such as variable based method, statistical indices, information theoretic, goodness of fit, etc.***

**Keywords:** Clustering, Hierarchical Clustering, Partitioning clustering

## 1. Introduction

Clustering is an important tool for a variety of applications in data mining, statistical data analysis, data compression, vector quantization, etc. Clustering is a process of grouping objects on the basis of similarities and dissimilarities between them. The main advantage of clustering over classification is that it is applicable to changes and helps single out useful features that distinguish different groups.

Clustering is one of the important techniques used in data mining. Data mining refers to the process of analysing the data from different perspectives and summarizing it to obtain useful information. Data mining allows users to analyze data from many different sources, classify and categorize it, and the relationships identified are summarized. Basically, data mining is the process of finding similarities or patterns among vastly different fields in large relational databases.

Any cluster should exhibit two main properties: low inter-class similarity and high intra-class similarity. Clustering algorithms are divided into two categories namely, hierarchical clustering and partitioning clustering. A hierarchical clustering algorithm is used to divide the selected data set into smaller subsets of data in hierarchical fashion. A partition clustering algorithm partitions the dataset into desired number of sets in a single step. K means is a well-known partitioning clustering technique that attempts to find a specific number of clusters (k), which are represented by their centroids.

The K-means clustering algorithm is as follows:

1) Select the initial centres of the K clusters. Then, repeat the step 2 followed by step 3 until the whole group of cluster stabilizes.
2) Generate a new partition by assigning each data to its nearest cluster centres.
3) Calculate the recently discovered cluster centres as the centroids of the clusters.

K-means may appear to be simple and applicable to a wide variety of data types. However it is sensitive to initial positions of cluster centres. Different initial centres may lead to different final clusters. The resulting clusters may not be optimal as the algorithm may converge to a local optimum. In order to reduce this effect, K-means can be run multiple times. There is also a possibility that an empty cluster may be formed if no points are assigned to it during the assignment step. Thus it is very important for K-means to have good initial cluster centres for good final clusters.

## 2. Related Work

K-means is a popular clustering algorithm. It aims at partitioning n observations into k clusters in which each observation belongs to the cluster of its nearest mean.

K-value is dependent on the characteristics of the dataset and may change over different datasets. The K-means clustering algorithm is simple and efficient clustering algorithm. It has a complexity of O(ntk) where n is the size of data set, t is the number of iterations and k is the number of clusters. As k value increases, number of iterations also increases.

In Data Analytics we often have very large data, which are almost similar to each other and hence we may want to organize them completely in a few clusters with same kind of observations within each cluster of its own. For example, in the case of customer data, even if we may have large amount of data from millions of customers, these customers may only belong to a few segments irrespective of the large amount of segments: customers may be almost alike within each segment but it may be different across the other segments. We may often want to analyse each segment separately, as they may behave differently (e.g. different market segments may have different product preferences and behavioural patterns).

In such type of situations, in order to identify such segments in the data, one should use numerical data techniques vastly

called Clustering techniques. Depended on how we define "alikeness" and "differences" between the observations made on data (e.g. customers or the customer assets), which can also be defined numerically using distance metrics, one can find various segmentation results.

Clustering techniques are mainly used to form different groups for data or their observations in a few segments so that data within such segment are alike while data across segments are different. Defining what we mean when we say "alike" or "different" observations is an important part of clustering analysis which often needs a lot of provisional information and creativity beyond what statistical tools can provide.

In the past various techniques have been implemented to determine the number of clusters for a given dataset. These techniques have been used for static as well as dynamic environments. In case of dynamic environments as the data changes, there is also a change in the mining algorithm used. The parameters for clustering also need to be modified. Here, partition clustering algorithm is used where the number of clusters need to be given prior to the execution. If the data is uncertain then, the number of clusters is executed at runtime. However this method is complex for execution and thus by executing the required data the complexity can be reduced.

Another clustering technique was applied in the field of image processing. The formation of various clusters for an image will help in analysis of the image. The hope is that the number of clusters within an image will be determined automatically. From the results of experimentation, it is not possible to identify definitively an index which will work well in all cases and hence be the most suitable as general index for all cases. However, in this case, the initial investigation should be expanded further such as further testing on different real data sets from a wide range of applications including medical images and reverse engineered data. Further tests should be carried out on clustering improvement on fuzzy methods using a wide range of validity index for automatic clustering algorithms.

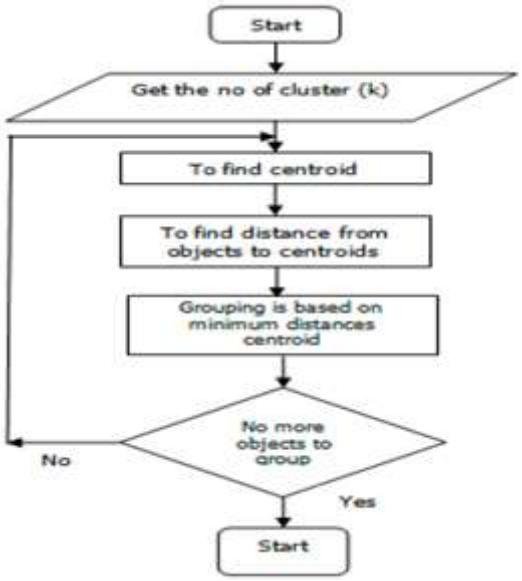
A new method to determine the correctness of clusters has also been implemented. This method uses k-means clustering. However, research is required to verify the capability of the method when applied to datasets with more complex object distributions. This method may also be expensive if used for large datasets.

Similarly, various other methods have been implemented such as using multi-layer clustering approach to determine the clusters. It provides a powerful tool for clustering complicated data. When applied to large scale data, successful implementation of multi-layer can be expected. There is another method called as AMOC (Automatic Clustering Technique) which inputs a large value of 'k' and then gradually decreases to find optimum clusters. However, this method is space consuming. An automatic clustering technique has been used in the TLBO framework. However, the major drawback of this method is that there may be formation of empty clusters, i.e. clusters which contain no element. This may result in the formation of suboptimal solutions.

Thus from the above surveys, it can be inferred that many methods and techniques have been suggested to determine the number of clusters automatically. However, they have certain drawbacks such as the number of clusters is taken as input from the user who is unaware of the dataset, there may be formation of empty clusters, the clusters may not be as per the requirement, the clusters formed may not be applicable for dynamic or changing data set, etc.

## 3. Proposed Work

The K-means algorithm is one of the renowned data clustering algorithms. In order to use K-means algorithm, it requires the number of clusters in the data to be pre-specified. Traditionally, finding the appropriate number of clusters for a given data set was generally a trial-and-error process. This project proposes a method based on information obtained during the K-means clustering operation itself to select the number of clusters, K. Also focuses to suggest suitable values for K, thus avoiding the need for trial and error.




The main goal of this project will be to develop a method to find the optimal number of clusters. The similarity/dissimilarity is measured in terms of distance function: d(i,j)It can be Manhattan distance or Euclidean distance.

1)Euclidean distance

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

:

2)Manhattan distance:

$$\sum_{i=1}^{k} | x_i - y_i |$$

## 4. Algorithm

The basic steps that we will perform are as follows:
Do the following steps for the no. of clusters (k) = 2 to (n-1)/2, where n is the size of the whole data set.

Step 1: Operate the basic fuzzy-c means clustering with k no. of clusters.
Step 2: Calculate the intra cluster similarity
Step 3: Calculate inter cluster similarity to other clusters
Step 4: Find the compactness distance
Step 5: Find the separation distance
Step 6: Calculate centroid of each cluster
Step 7: Find optimal cluster number
Step 8: Display cluster results.

## 5. Conclusion and Future Work

The K-means clustering algorithm is one of the split clustering algorithms that initially depends on two factors which are the initial clusters and the k value. Primarily, in K-means clustering, initial clusters are based on the algorithm which randomly selects centroids and randomly chooses k values.

As the data changes, the data mining algorithm will also have to adapt to these changes and it has to modify its specifications. In these cases where the user has to give the number of clusters before its execution, we select this partition clustering algorithm. If data is not certain in some cases, the number of clusters can be decided at run time. In such cases, we proposed a method to find the ideal number of clusters for a given dynamic data set and display the observations.

## References

[1] Chatti Subbalakshmi, G Rama Krishna, S Krishna Mohan Rao, P Venketeswa Rao,"A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set", International Conference on Information and Communication Technologies (ICICT 2014)
[2] E.A. Zanaty,Determining the number of clusters for kernelized fuzzy C-means algorithms for automatic medical image segmentation, Egyptian Informatics Journal (2012)
[3] D T Pham, S S Dimov, and C D Nguyen," Selection of K in K-means clustering", September 2004
[4] Mingjin Yan," Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion", November 2005
[5] K. Karteeka Pavan, Allam Appa Rao, A.V. Dattatreya Rao, "An Automatic Clustering Technique for Optimal Clusters"
[6] Boris Mirkin," Choosing the number of clusters"
[7] Rashmi Gangadharaiah, Ralf D. Brown, Jaime Carbonell," Automatic Determination of Number of clusters for creating templates in Example-Based Machine Translation"
[8] Ramachandra Rao Kurada, Karteeka Pavan Kanadam," A generalized automatic clustering algorithm using improved TLBO framework",2015
[9] David Sebiskveradze, ValeriuVrabie, Cyril Gobinet, Anne Durlach, Philippe Bernard, Elodie Ly,MichelManfait, Pierre Jeannesson and Olivier Piot," Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections", 2011
[10] Trupti M. Kodinariya, Dr. Prashant R. Makwana,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