ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 7, July 2018

# Comparison of various clustering algorithm with proposed clustering algorithm

[1] RadhikaSethi, [2] Sikander Singh Cheema
[1][2] Department of Computer Engineering, Punjabi University, Patiala, India.

**Abstract: Division of data into similar groups of objects called clustering. Each group is called cluster. A comparison between all the clustering algorithms i.e. K-means, Exception Maximization, Hierarchical, Density Based, Farthest First, SOM are thought about on the bases of size of informational collection, number of clusters and time taken to Shape groups.**

**Keywords: Clustering, K-means , Expectation Maximization, Hierarchical, Density clustering, Algorithm**

## INTRODUCTION

Clustering is division of information into gathering of comparable items, each gathering is called cluster, comprise of different articles that are comparative among themselves and unique contrasted with protest of different gatherings. Representing data by fewer clusters necessarily loses certain fine details, but achieve achieves simplification. It represents many data objects by few clusters, and hence, it models for data by its cluster. Clustering analysis is the organization of a collection of patterns into cluster based on similarity. Patterns with in valid clusters are more similar to each other then they are to a pattern belonging to a different cluster. Some researchers improved some data clustering algorithms, other implemented new ones, and some studied and compared different clustering algorithms. The experiments were conducted on the WEKA (3.7.10) tool to compare different clustering algorithm. it was used because it provides better interference to the user than compare to other data mining tools. Clustering Algorithms which are compared are partitioning based i.e. K- means, Farthest First, Expectation Maximization, and Non Partitioning Based i.e. Density based, Hierarchal Based, Cobweb.

### *A. K-means Clustering Algorithm*

K-means clustering Algorithm is First Proposed by Macqueen in 1967 which was uncomplicated, non supervised learning clustering algorithm. K-means is a dividing grouping calculation, this procedure is utilized to arrange the given information objects into k different clusters through the iterative strategy, which has a tendency to merge to a neighborhood least. So the result of generated clusters is dense and independent of each other. The algorithm comprises of two distinct stages.. These two steps are rehashed till the inside cluster variety can't be diminished any further. The inside cluster variety is figured as the aggregate of the Euclidean separation between the information focuses and their particular cluster centroids

Euclidean Distance = $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

In the primary stage user chooses k centeroid haphazardly, where the value of k is settled ahead of time. To take each data object to the nearest centre. A wide range of Distance functions are considered to determine the distance between each data object and cluster centers. At the point when every one of the information objects are incorporated into a few bunches, the initial step is finished and early gathering is finished. At that point the second stage is to recalculate the normal of early framed groups. This iterative Procedure proceeds over and again until the criterion function becomes the minimum

### *B. Expectation Maximization Clustering Algorithm*

Expectation Maximization (EM) algorithm is an iterative method for finding Maximum likelihood or maximum an osterioro (MAP) estimates of parameters in Statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an exception (E) step, which creates a function for the parameters, and a maximization (M) step, which processes parameters maximizing the expected log-likelihood found on the E step. These parameters-estimates are then used to decide the dispersion of the idle factors in the following E step. EM assigns appoints a likelihood dispersion to each instance which shows its likelihood having a place with each other clusters. The EM (expectation maximization) technique is similar to the K-Means technique. The basic operation of K-Means

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 7, July 2018

clustering algorithms is relatively simple: Given a fixed number of k clusters, assign observations to those clusters so that the means across clusters (for all variables) are as different from each other as possible. The EM algorithm extends this basic approach to clustering in two important ways:

Instead of assigning examples to clusters to maximize the differences in means for continuous variables, the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters.

Expectation Maximization algorithm the basic approach and logic of this clustering method is as follows. Suppose you measure a single continuous variable in a large sample of observations. Further, suppose that the sample consists of two clusters of observations with different means (and perhaps different standard deviations); within each sample, the distribution of values for the continuous variable follows the normal distribution. The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data (distribution). Put another way, the EM algorithm attempts to approximate the observed distributions of values based on mixtures of different distributions in different clusters. The results of EM clustering are different from those computed by k-means clustering. The latter will assign observations to clusters to maximize the distances between clusters. The EM algorithm does not compute actual assignments of observations to clusters, but classification probabilities.

***C. Farthest First Clustering Algorithm***

Farthest First is a modified K-means that places each cluster center in turn at the point further most from the existing cluster center. This point exists within the data area. This greatly increases the clustering speed in the vast majority of the cases since less reassignment Furthermore, adjustment is required.

***D. Hierarchal Clustering Algorithm***

Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. In Contrast, hierarchical algorithms combine or on the other hand separate existing groups, making a Various leveled structure that mirrors the request in which clusters are grouped or isolated. In an agglomerative method, Which builds the hierarchy by merging, the objects initially belongs to a list of singleton sets s1,..,S2,Sn. Then a cost function is used to find the pair of sets {si,sj} from the list that is the “cheapest” to merge. Once merged Si and Sj are removed from the list of sets and replaced with si U sj. This process iterates until all objects are in single group. Different variations and agglomerative hierarchical clustering algorithm may use different cost functions. Complete Linkage, average Linkage and single linkage methods are maximum, average and minimum distances between the members of two clusters respectively.

Hierarchal strategy makes a hierarchal disintegration of the given arrangement of information objects shaping a dendogram-a tree which parts the database recursively into littler subsets. The dendogram can be shaped in two different ways bottom up and top down. Hierarchal algorithm combines or on the other hand partitions existing groups, making a hierarchal structure that mirrors the request in which clusters are combined or separated. The base up approach, likewise called the "agglomerative" approach, begins with each question framing a different group. It progressively combines the articles or groups as per a few measures like the separation between two focuses of two groups and this is finished until the point that the greater part of the groups are converted into one, or until an end condition holds. The best down likewise called "devise approach", begins with every one of the items in a similar cluster. In each Progressive cycle a group is part into small clustering like manner to a few measures until the point when in the end each object is in one cluster, or until the point when an end condition holds. Following is the Pseudo code of the hierarchical clustering algorithm to explain how it works:

- Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.
- Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.
- If all patterns are in one cluster, stop. Otherwise go to step 2.

The advantages of hierarchical clustering algorithms are the reason this algorithm was chosen for discussion.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 7, July 2018

- Embedded flexibility regarding a level of granularity.
- Ease of handling of any forms of similarity or distance.
- Consequently applicability to any attributes types.
- Hierarchical clustering algorithms are more versatile.

***E. Density based Clustering Algorithm***

Density based clustering algorithm attempt to discover clusters on the bases of data points in a region. The key thought of density based clustering is that for each occurrence of a group the area of a given span needs to contain at any rate least number of occurrences.try to find clusters based on density of data points in a region.

***F.Self -Organization Map Algorithm***

Self organization Map(SOM) uses a competition and cooperative mechanism to achieve unsupervised learning. In the classical SOM, a set of nodes is arranged in a geometric pattern typically 2 dimensional lattice.

According to Osama Abu Abas compared the performance of different Four clustering algorithms according to the factors

- Size of data set.
- Number of data set
- Number of cluster
- Type of software

For each factor, four tests are made, one for each algorithm. For example according to the size of data, each of the four algorithms: k-means, Hierarchal Clustering, SOM and EM is executed twice. First by trying huge data set ,then by small data set. The total number of times The algorithm Have been executed is 32. Four each 8-runs group, the result of executions is studied and compared.

Hierarchal clustering was compared with other algorithms; the hierarchal tree is cut at two different levels to obtain corresponding number of clusters. As a result the value of k becomes greater the performance of SOM becomes lower. The performance of K-means and EM algorithms become well than hierarchal clustering algorithm.

***Table 1. The relationship between number of clusters and the performance of algorithms.***

| Number of clusters | Performance | | | |
|---|---|---|---|---|
| | SOM | K-means | EM | HCA |
| 10 | 58 | 60 | 59 | 63 |
| 14 | 65 | 70 | 68 | 67 |
| 08 | 73 | 76 | 75 | 78 |
| 56 | 82 | 85 | 85 | 88 |

According to the accuracy SOM shows more accuracy in classifying most objects to their clusters than other algorithms, but the number of K becomes greater the accuracy of hierarchal clustering becomes better until it reaches the accuracy of SOM algorithm.

According to the size of data set, the quality of EM and K-means algorithms becomes very good when using a huge dataset.

According to the type of data set when random and ideal data sets are used As a result Hierarchal clustering and SOM algorithms give better results than when using random data set, But when ideal data sets were used K-means gives better result. it indicates that k-means and EM are very sensitive for noise in the dataset.

According to the type of the software, two packages were used to compare between the algorithms. Running the clustering algorithms using any one of them gives almost same results even when changing any of the three factors (data type, data set and number of clusters)

## CONCLUSION

- As the number of clusters, becomes greater the performance of SOM algorithm becomes lower.
- The performance of K-means and EM algorithm is better than hierarchal clustering algorithm.

- As the value of K-means becomes greater, the accuracy of Hierarchal clustering becomes better until it reaches the accuracy of SOM algorithm.
- The quality of EM and K-means algorithms become very good when using huge data set.
- As a general conclusion Partition of algorithms like k-means and EM are recommended for huge data set while hierarchal clustering algorithms are recommended for small dataset.

## REFERENCES

[1] G. Sehgal, Dr. K. Garg, Comparison of various clustering, International Journal Of computer Science and Information Technologies, Vol.5 ( 3) 2014, 3074-3076.

[2] O.A.Abbas, Comparison Between Data Clustering Algorithms, International Arab Journal of Information Technology, Vol5, No3, July 2008

[3] Han J. and Kamber M., Data mining Concepts and Techniques, Morgan Kaufmann Publishers 2001.

[4] N.Sharma, A.Bajpai, R.Litoria, Comaprison the various clustering algorithms of weka tools, International Journal of emerging technology and Advanced Engineering, Volume 2, Issue 5 , ISSN: 2250-2459 May 2012.

[5] M. Eisen , Cluster and Tree View Manual,Stanford University, 1998.

[6] A,. Jain, M.Murty, P.Flynn Data Clustering : A Rivew, ACM Computing Surveys, Vol.31,no 3,1999.

[7] A.Riabov, Z.Liu, L.Zhang, Clustering algorithm for content based publication-subscription systems In Proceedings of the 22nd International Conference on Distributed Computer Systems USA, pp 133,2002.

[8] G. Chen, S.Jaradat, N.Banerjee, T.Tanaka, M. Jhang, Evaluation and comparison of Clustering Algorithm in Analyzing ES Cell Gen Expression data, Satistica Sinca, Vol.12,PP.241-262,2002.

[9] H.JHA, C.Ding, H. Simon, Biparatite Graph Partiotioning and Data Clustering in proceedings of 10th International Conference on Information and Knowledge Management, ACM press,PP.25-32.

[10] M. Montes Y-G Mez M., A. Gelbukh, and A. Lpez, Text Mining at Detai Level Using Conceptual Graphs, Lecture Notes in Computer Science Vol. 2393, PP 122-136,2002.