Volume 3, Issue 5, May 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Hybrid Clustering Algorithm with Modifications Enhanced K-Means and Hierarchal Clustering

**Gurjit Singh***
*Department of Computer science & Engineering*
*Sri Guru Granth Sahib World University,*
*Fatehgarh Sahib, Punjab, India.*

**Navjot Kaur**
*Department of Computer science & Engineering*
*Sri Guru Granth Sahib World University,*
*Fatehgarh Sahib, Punjab, India.*

***Abstract*— Clustering is an essential task in Data Mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. K-Means clustering is a clustering method in which the given data set is divided into K number of clusters. This paper is intended to give the introduction about K-means clustering and its algorithm. The experimental result of K-means clustering and its performance in case of execution time is discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution, time we are using the Ranking Method. And also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using the Ranking Method. Modifications in hard K-means algorithm such that algorithm can be used for clustering data with categorical attributes. to use the algorithm for categorical data modifications in distance and prototype calculation are proposed. To use the algorithm on numerical attribute values, means is calculated to represent centre, and Euclidean distance is used to calculate distance.***

***Keywords*— Clustering, Hierarchical Clustering, K-means, Ranking method, SOM**

## I. Introduction

In today's highly competitive business environment Clustering play an important role. As K- means Clustering is a method for making groups of the data set or the objects that are having similar and dissimilar properties. how in K-means algorithm the distance between the objects and mean is calculated and the methods of selecting initial points in K-means Clustering algorithm then contains main steps in K-means clustering algorithm "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points" In this paper the uniform distribution of the data points is discussed that how this approach reduce the time complexity of the K-means clustering algorithm. By using this approach the elapsed time is reduced and the cluster is of better quality. In this a very good method is used for finding the initial centroid. In this initially, the distance between each data points is computed. Where *X* denotes the original data set, *Ci*, *Cj* are clusters of X, and n is the number of clusters

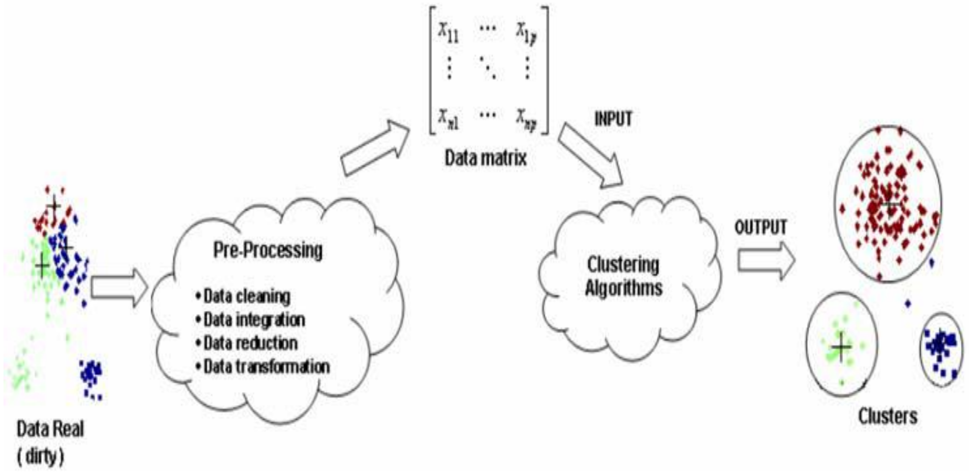

**Fig. 1 clustering process**

## II. K-MEANS CLUSTERING ALGORITHM

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The result of partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.

Types of Clustering Algorithms are:
1. K-means Clustering Algorithm
2. Hierarchical Clustering Algorithm
3. Density Based Clustering Algorithm
4. Self-organization maps (SOM)
5. EM clustering Algorithm

### STEPS OF K-MEANS CLUSTERING ALGORITHM

K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters; the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroid of each cluster is selected for clustering and then according to the chosen centriod, the data points having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroid. This algorithm consists of four steps:

1. Initialization: In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification: The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. Centroid Recalculation: Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. Convergence Condition: Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

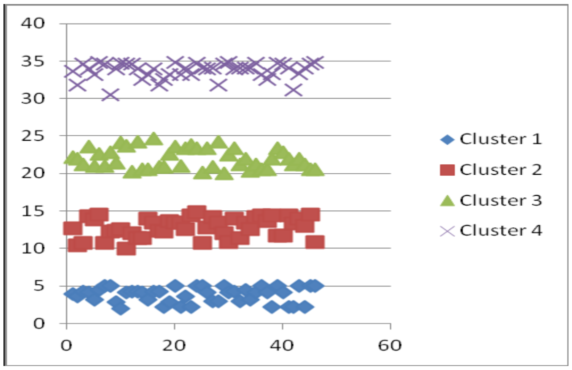

**Fig.2 Graph Shows Clusters in K-means Clustering Algorithm**

**Main advantages:**

1. K-means clustering is very Fast, robust and easily understandable. If the data set is well separated from each other data set, then it gives best results.

2. The clusters do not having overlapping character and are also non-hierarchical in nature.

**Main disadvantages:**

1. In this algorithm, complexity is more as compared to others.

2. Need of predefined cluster centers.

3. Handling any of empty Clusters: One more problems with K-means clustering is that empty clusters are generated during execution, if in case no data points are allocated to a cluster under consideration during the assignment phase.

The experimental results demonstrated that the proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

### K-MEANS ALGORITHMS

*1.1 Hard C- Means (HCM):* In HCM each object is assigned to exactly one cluster.

*1.2 Fuzzy C-Means (FCM):* FCM allows one data object to belong to two or more clusters at the same time.

FCM becomes very sensitive to noise and outliers because data point memberships are inversely related to the relative distance of the data to the cluster centers.

*1.3 Rough C-Means (RCM):*The rough set is a mathematical tool for managing uncertainty that arises from the indiscernibility between objects in a set.

*1.4 Rough, Fuzzy, Possibilistic C-Means (RFPCM)*: RFPCM adds both probabilistic and possibilistic memberships and the lower and upper approximations of rough sets into *c*-means algorithm**.**

HCM, FCM, RCM to RFPCM. So we can say RFPCM for categorical data performs better over other c-mean variants. Among these algorithms RFPCM gives improved results over other variations of k-means algorithm generated by this system is less. The major issues concerning data mining in large databases are efficiency and scalability. While in case of high dimensional data, feature selection is the technique for removing irrelevant data. It reduces the attribute space of a feature set. More reliable estimation of prediction is done by f-fold –cross- validation. The error rate of a classifier produced from all the cases is estimated as the ratio of the total number of errors on the hold-out cases to the total number of cases. By increasing the model complexity, accuracy of the classification is increases. Over fitting is again major problem of decision tree. The system has also facility to do post pruning that is through reduced error pruning technique. Using this proposed system Accuracy is gained and classification error rate is reduced compare to the existing system.

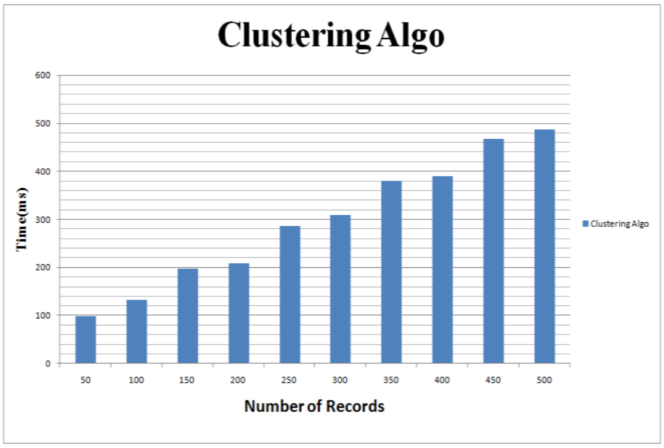

**Fig.3 Execution time of K-means clustering algorithm.**

.

## III. Clustering Algorithms

In this section clustering algorithms divided into the following major categories [3]:

*1. Sequential algorithms:* These algorithms produce a single clustering. They are quite straightforward and fast methods. In most of them, all the feature vectors are presented to the algorithm once or a few times (typically no more than five or six times). The final result is, usually, dependent on the order in which the vectors are presented to the algorithm.

2. *Hierarchical clustering algorithms:* These schemes are further divided:

*2.1 Agglomerative algorithms (bottom-up, merging):* These algorithms produce a sequence of clustering of decreasing number of clusters, *m*, at each step. The clustering produced at each step results from the previous one by merging two clusters into one. The main advantage of HAC is the user can guess the right partitioning by visualizing the tree, he usually prune the tree between nodes presenting an important variation. The main disadvantage is that requires the computation of distances between each example, which is very time consuming when the dataset size increases.

*2.2 Divisive algorithms (top-down, splitting):* These algorithms act in the opposite direction; that is, they produce a sequence of clustering of increasing *m* at each step. The Clustering produced at each step results from the previous one by splitting a single cluster into two.

*2.3 Clustering algorithms based on cost function optimization:* This category contains algorithms in which "sensible" is quantified by a cost function, *J*, in terms of which a clustering is evaluated. Usually, the number of clusters *m* is kept fixed. Most of these algorithms use differential calculus concepts a produce successive clustering while trying to optimize *J*.

Algorithms of this category are also called iterative function optimization schemes. This category includes: Hardor crisp clustering algorithms, Probabilistic clustering algorithms.

## IV. An Applied Methodology For Clustering Process

Today SOM is an advanced analytical tool applied in exploratory data analysis. As an unsupervised neural network it shows a great potential and power in the domain of visualizing multidimensional data sets and data clustering. In contrast to the standard clustering methods, such as K-means, self-organizing maps provide easy visualization and impose few assumptions and restrictions. In the neural network model of the SOM there are two layers: input layer of units and output layer given in the form of two-dimensional grid of neurons. The relations between data items become explicit in the SOM due to a nonlinear projection from a high-dimensional space onto a two-dimensional display. The SOM map may be presented as an "elastic net" that is stretched to cover the input space of data. And additionally, component planes which usually accompany the map provide information about data division in the input space, relative distribution of the components (variables) and the visual image of data correlations.

The model owes its popularity to its capability of presenting on the output layer a very intuitive description of the similarity among groups of data in the input space. Namely, the SOM neural network represents the image of the whole observation space: different neurons on the map are representatives of the different observation domains.

*3.1 CHAID decision tree model:* The CHAID algorithms consist of three steps: merging, splitting and stopping. A tree is grown by repeatedly using these three steps on each node starting from the root node. At each step, CHAID chooses the independent variable that has the strongest interaction with the dependent variable. The main statistic used in this algorithm is *chi-square* statistic, which is accompanied by respective *p*-value. Categories of each independent variable are merged if they are not significantly different with respect to the dependent variable. The "best" split for each independent variable is found in the merging step. For the merging and splitting purpose the adjusted *p*-value of *chi-square* statistic is calculated for each independent variable. The selection of independent variable is determined by the smallest adjusted *p*-value (i.e., most significant). If it is less than *a priori* specified α-level the relevant tree node is split by this independent variable. Also, during the tree growing process the stopping step checks if the process should be stopped according the precisely defined stopping rules.

*3.2 The combined approach of two models:* It is pointed out in the introduction section that this paper proposes the combined approach of Kohonen SOM model and CHAID decision tree model as a two-stage clustering strategy. At the first stage the SOM algorithm is used. It provides an apparent visualization of the data structure with an indicative clustering solution. But, finding clusters is not an end in itself. It should be followed by investigation of the clusters' meaning. The good understanding of data structure is usually connected with the knowledge on mutual interactions of independent variables with the concrete clustering variable. In this respect the SOM component planes may be very informative and helpful. But, they can be improved by the usage of a supervised clustering technique that can import the clustering result from the SOM map and afterwards explore the input-output variable relations in a more detailed and visually transparent way.

For this purpose the application of CHAID decision tree model (as a second stage of clustering process) is proposed. It gives a good explanation of input-output relationships between data variables in a visually very attractive and easy-to-understand decision tree form. From the methodological point of view there is a fundamental difference between Kohonen SOM, on the one hand, and decision tree model, on the other. The former model is an unsupervised clustering model: the training (or learning) process is running in the absence of an output variable acting as a supervisor. In contrast to this, in the case of CHAID model the clustering is supervised, that is, measured against a response variable whose values are known (5). In the combined approach the clustering variable is generated from the SOM network and further transferred in the CHAID model as the dependent variable. Concerning the input variables (i.e. predictors in CHAID Terminology) the same variables are used in both modelsocess should be stopped according the precisely defined stopping rules

## V. Future Work

In future, in case of clustering the marks of students from different-2 databases are considered by using the concept of Query redirection. By using the Query redirection approach we can easily cluster the large amount of data from distributed environment as from different databases. So if this approach is considered, then the performance of K-means clustering algorithm is improved for large samples of data set that are also distributed in nature.

## VI. Conclusion

Modified k-means algorithm gives reduced value of objective function for categorical data clustering. If we observe stability of algorithm in terms of objective function value for minimum value and converged value, these values are equal or almost equal. Results show that there is significant reduction in objective function value from maximum (which occur at first iteration) to local minimum or converged value of objective function for each algorithm whereas values are decreasing in sequences from HCM, FCM, RCM to RFPCM. So we can say RFPCM for categorical data performs better over other c-mean variants. Among these algorithms RFPCM gives improved results over other variations of k-means algorithm.

This paper discusses the combined application of two clustering methodologies: Self-organizing map (SOM) and CHAID decision tree model as parts of a two-step clustering strategy. The first step of the strategy is focused at the SOM application which provides an adequate clustering solution and the optimal cluster number. The second step concerns to the CHAID analysis of the previously generated clustering results. The proposed clustering strategy is applied on the concrete example concerning the market segmentation process. It is empirically demonstrated that Kohonen SOM methodology can be successfully supplemented with the CHAID decision tree model. While component planes of SOM map are providing the visual image of the individual contribution of the respective input variables to the final cluster structure, the CHAID decision tree reveals the variable interactions within concrete clustering solution. That way CHAID decision tree adds an appropriate visual touch to the original SOM model. Applying on the marketing data we have extracted the knowledge: number of segments, their structure, and individual and combined contributions of relevant input variables to the clustering process and the interdependence of the input variables. All this information can be a starting point for better understanding the relevant customer habits and, also, an adequate setting for the target marketing promotions.

**References**

[1]. Jian Yu," General C-Means Clustering Model", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, VOL. 27, NO. 8, PP.1197-2111, 2005.

[2]. Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li," Automated Variable Weighting in k-Means Type Clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* VOL. 27, NO. 5, PP. 657-668, 2005.

[3]. Carlos Ordonez ," Integrating K-Means Clustering with a Relational DBMS Using SQL", *IEEE Trans. Knowl. Data Eng.,*, VOL. 18, NO. 2, PP. 188-201, 2006.

[4]. Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, and Andre C. Ponce Leon F. de Carvalho ," A Survey of Evolutionary Algorithms for Clustering", *IEEE Trans. Syst., Man, Cybern.—Part C: Appl. And Review,* Vol. 39, No. 2,PP.133-155,2009

[5] P. Maji and S. K. Pal, "Rough–fuzzy C-medoids algorithm and selection of bio-basis for amino acid sequence analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 6, pp. 859–872, Jun. 2007.

[6] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek," A Possibilistic Fuzzy c-Means Clustering Algorithm," *IEEE Trans. Fuzzy Syst*, Vol. 13, no. 4, Aug 2005.

[7] S. Mitra, H. Banka, and W. Pedrycz, "Rough–fuzzy collaborative clustering," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 795–805, Aug. 2006.

[8] Zhexue Huang, Michael K. Ng.," A fuzzy k-modes algorithm for clustering categorical data," *IEEE Trans. on fuzzy systems.*, Vol 7 No 4. August 1999.

[9] Chen Ning, Chen An, Zhou Long-xiang ," Fuzzy k-prototypes algorithm for clustering mixed Numeric and categorical valued data," *Journal of software* Vol.12 No. 8,2001.

[10] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, May 1993.

[11] Pawan Lingras, Min Chen, and Duoqian Miao," Rough Cluster Quality Index Based on Decision Theory," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 7, July 2009.

[12] Tapas Kanungo,David M. Mount, Nathan S. Netanyahu,Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, VOL. 24, NO. 7, PP. 881-892, 2002.