Hierarchical k-Means Algorithm(hk-Means) with Automatically Detected Initial Centroids

Vaishali R. Patel
Department of Computer Engineering
Shri S'ad Vidya Mandal Institute of Technology
Bharuch, Gujatat, India
vaishalirajpatel@gmail.com

Rupa G. Mehta
Department of Computer Engineering
Sardar Vallabhbhai National Institute of Technology
Surat, Gujarat, India
rgm@coed.svnit.ac.in

***Abstract*— Unsupervised learning is a technique to organize the data into meaningful way having similarity. Cluster analysis is the study of clustering techniques and algorithms which are helpful to discover important patterns from fundamental data without knowledge of category label for further analysis. k-Means algorithm is one of the most popular clustering algorithm among all partition based clustering algorithm to partition a dataset into meaningful patterns. k-Means algorithm suffers from the problem of specifying the number of clusters in advance and often converges to local minima and therefore resulted clusters are heavily dependent on initial centroids. Various methods have been proposed for automatic detection of initial centroids to improve the performance and efficiency of k-Means algorithm. This paper presents an overview of clustering, clustering techniques and algorithms, addressing problems of k-Means algorithm, comparison of different methods for automatic detection of initial centroids and propose a new Hierarchical method(hk-Means) for automatically detection of initial centroids in k-Means algorithm, implementation of traditional k-Means with automatic pre-process dataset to remove noise.**

***Keywords*— Clustering, Initial Centroids, k-Means Preprocessing, Outlier**

I. INTRODUCTION

Data Mining is the process to extract previously unknown useful information from huge collection of data. Goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of "natural," hidden data structures, rather than provide an accurate characterization of unobserved samples generated from the same probability distribution[1]. People always learn new object and properties of that object which can be comparable with other objects based on similarity or dissimilarity. Clustering is an unsupervised learning technique(learning without an instructor) of data mining where huge collection of data are classified into clusters having similarity among data objects. Various clustering algorithms according to different techniques have been designed and applied to various data mining problems successfully. These algorithms promise to generate efficient and quality clusters, but the quality of clustering techniques depends on; similarity measure and technique used for implementation, ability to discover some or all hidden patterns, definition and representation of cluster chosen[2]. Accomplishment of clustering algorithms in many areas, it causes many precincts to the researchers when no or little information are available. There is also no universal clustering algorithm developed; that's why it is very crucial job to choose appropriate clustering technique and algorithm considering above precincts. A simple and commonly used algorithm for producing clusters by optimizing a criterion function, defined either globally (over all patterns) or locally (on a subset of the patterns), is the k-means algorithm[5]. The k-Means algorithm[3][4][5][6][7] is effective in producing clusters for many practical applications. This algorithm results in different types of clusters depending on the random choice of initial centroids. Several attempts were made by researchers to improve the performance of the k-means clustering algorithm. In this paper, we have study and provide an overview of cluster analysis, clustering techniques and algorithms, addresses problems with traditional k-Means clustering algorithm. This paper also compare various methods proposed by many researchers to make an improvement in traditional k-Means clustering algorithm for the automatic detection of initial centroids with pros and cons. To make enhancement in the earlier proposed methods to improve traditional k-Means, we propose a new hierarchical k-Means method to automatically detect initial centroids with preprocessing task. This paper presents the implementation of traditional k-Means algorithm with preprocessing task to remove noise and result on RIVER dataset.

II. CLUSTERING AND ALGORITHMS

Clustering is important for automatically organization of data coming from various data sources. Classically, these data are described as a set of objects characterized by a set of features. Cluster analysis is the organization of a collection of objects which are represented as vector of measurements into clusters based on similarity. Objects in cluster are similar to each other than the objects belonging to a different clusters. It is important to understand the difference between unsupervised classification (clustering) and supervised classification. In supervised classification, we are having labeled objects. Problem with supervised classification is to label a new unlabeled object. Training objects gain

knowledge of details of classes and it is then used to label a new object. Problem with unsupervised classification(clustering) is to cluster a given collection of unlabeled objects into significant clusters. From a practical perspective clustering plays an outstanding role in data mining applications. Sometimes cluster analysis requires to execute number of components repeatedly; because there is no universal technique for cluster analysis process. There is a close relationship between clustering techniques and many other disciplines; it has been applied in a wide variety of fields like engineering, economics computer, social and medical sciences, pattern recognition; compression; classification; and classic disciplines as psychology and business. Figure 1 illustrates cluster analysis process having six stages discussed below.

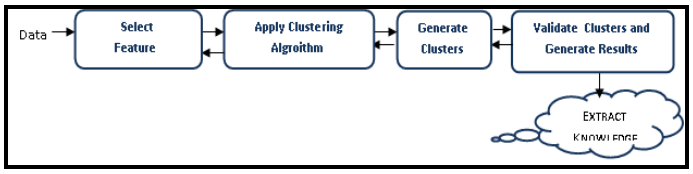

Figure-1: Steps of Cluster Analysis Process[8]

1) *Input Data*: Fundamental data from data source are chosen as input and passed to second stage for analysis.

2) *Select or Extract Feature*: It selects an essential feature of an object which differentiate it from other objects and extracts most promising or best feature[9] useful to generate new features. Clustering process can be simplified by proper selection of features.

3) *Apply Clustering Algorithm:* During this stage, selection of corresponding proximity measure and criterion function is constructed. "It has been very difficult to develop a unified framework for reasoning about clustering at a technical level, and profoundly diverse approaches to clustering" [10]. It is important to select the proper feature of the problem to design or apply clustering algorithm.

4) *Generate Clusters:* During this stage, number of clusters are generated after applying appropriate clustering algorithm on given fundamental data.

5) *Validate Clusters and Generate Results:* During this stage, different validation criteria based on applied clustering technique is used to check the significance of generated clusters and then generate results. Appropriate selection of criterion is still problematic and require more efforts.

6) *Extract Knowledge*: During this stage, useful knowledge is extracted from results. Number of experiments is done to guarantee that extracted knowledge is reliable for further processing.

A. Clustering Techniques and Algorithms Clustering problem is represented by different ways. Different Clustering algorithms are designed and apply to solve the clustering problem. This section summarize clustering algorithms. Clustering techniques are broadly divided into following.

a) *Partitioned Clustering*: This technique partition data into number of subsets where each subset represents a cluster and each cluster must contain at least on object and each object must belongs to exactly one cluster. Well known partitioning algorithms are:k-Means,k-Medoids,PAM,CLARA,CLARANS.

b) *Hierarchical Clustering*: Hierarchical clustering method having main two approaches: (i) Agglomerative (ii) Divisive. In Agglormative Approach; it generates a cluster by merging two smaller clusters in a bottom-up manner. So, generated clusters form a binary tree hence root node is a cluster having all data objects which are leaf nodes. Several agglormative clustering algorithms are BIRCH, CURE, ROCK and Chameleon In divisive approach; It splits a cluster into two smaller clusters in a top-down manner which constructs a binary tree. It repeatedly splits a current cluster until the number of clusters reaches a predefined value K, or some other stopping criteria are met [11]. There are two divisive clustering algorithms, named MONA and DIANA [12].

c) *Density Based Clustering:* This technique consider the density around each point to demonstrates boundaries and identify the core cluster points. The close cluster points in a single neighborhood are then merged. It can find clusters of arbitrary shape and handle noisy data. Many input parameters that are difficult to define Source. DBSCAN and OPTICS algorithms are examples of density based clustering technique.

d) *Grid Based Clustering*: Grid based clustering approach uses a multi resolution grid data structure. It quantizes the space into a finite number of cells to form a grid structure where clustering operations are performed. It is popular for its fast processing time. STING, CLIQUE and Wave Cluster are examples of this technique.

e) *Model Based Clustering:* This methods attempt to optimize the fit between the given data and some mathematical model. Some methods are often based on the assumptions. This method follow main two approaches: i) statistical ii) Neural Network. COBWEB is an example of statistical approach of model based technique.

III. TRADITIONAL K-MEANS ALGORITHM

In this section, we discuss the working of traditional k-Means clustering algorithm. K-Means algorithm is one of the most popular clustering algorithm due to its efficiency and simplicity in clustering large data sets. In traditional k-Means algorithm, a set of data set D is classified using a certain number of clusters (k clusters) which are initialized apriori. It define k centroids, one for each cluster and then consider data object belonging to the given data set and associate this data objects to the closest centroid. Euclidean distance generally considered to determine the distance between data objects and the centroids [13]. First step is completed when there is no data object is remaining and early group is done. Here, there is need to re-calculate new centroids. After obtaining new centroids same data objects are binded with the closest centroid and generate a loop. At the end of loop, k-centroids change their point step by step until centroids do not move any more. This algorithm works on basis of minimizing squared error function. The k-Means algorithm always converges to a local minimum. Local minimum found depends on the initial cluster centroids. Block diagram of traditional k-Means clustering algorithm is represented in Figure 2.

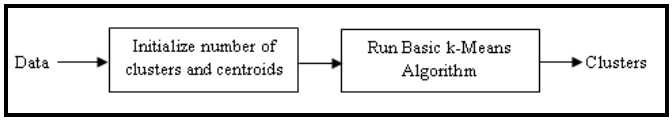

Figure 2: Block Diagram of Traditional k-Means Algorithm

Pseudo code for the traditional k-means clustering algorithm is listed in Figure 3:

Algorithm: Traditional **k-Means** – k-Means Clustering

Input: D is a set of n data objects.
k is number of desired clusters

Output: set of k clusters.

Method:

1. Arbitrarily choose k data-items from D as initial centroids;
2. Repeat

   Assign each item di to the cluster which has the closest centroid;

   Calculate new mean for each cluster;

   Until convergence criteria is met.

Figure 3: Traditional k-Means Algorithm[14]

IV. PROBLEMS WITH K-MEANS AND EXISTING METHODS

In this paper, we address problems with k-means algorithm. First, in k-Means, we need to determine number of cluster and initial centroids preliminary. Due to random selection of initial centroids, k-means often converges to local minima. So, the selection of initial centroid is very important. As there is no standard method in k-Means algorithm to select initial centroids, resulted clusters may be different. K-Means algorithm gives better results only when the initial partitions are close to the final solution [15]. Initial cluster centroids are either selected randomly or first n data objects are selected. Many researchers have proposed methods to make enhancement in traditional k-Means for automatic detection of initial centroids. In this section, we have analysed and presented the existing methods.

Fahim has proposed a method[17] to find good initial centroids by partitioning dataset into number of blocks and k-means is apply to each block. This method finds good initial centroids but it is complex and time complexity is more.

This proposed method[18] finds the better initial centroids. Dataset is first checked and convert negative value of attributes into positive. It then calculates the distance of each data object from the centroid and heuristic approach is used to assign each data object to initial centroid. Data objects are sorted according to the distance. Sorted data objects are partitioned of equal size and middle data object is considered as initial centroid with less time complexity. These methods do not work well for high dimensional data sets.

Zhong Wei et. al. has proposed method[21] uses greedy initialization approach to overcome potential problems of random initialization of centroids. In this method, clustering algorithm will be performed for number of iterations during each run. After each run, initial data objects are selected as cluster having good structural similarity and it's distance against all data objects already selected in the initialization array. If the minimum distance of the new data objects is greater than the specified distance, these data objects are added to the initialization array.

Tajunisha and Saravanan have proposed a method[16] that reduces the dimension and finds the initial centroids using PCA. Heuristics approach is used to reduce the number of distance calculation in the standard k-Means algorithm to improve the efficiency of this method.

Clustering accuracy can be improved[14] by finding the initial centroids for k-means algorithm and cosine measure is used to find the informative genes. This method uses the k-nearest neighbour (KNN) algorithm to fill missing values. This method identifies eigenvectors of corresponding largest eigenvalues for class partition and then identifies initial centroids. Cosine measure is used to compute the similarity among the data objects (genes).

The method proposed by J. Kleinberg[10] provides an algorithm that uses two methods: one to find initial centroids and second to assign data objects to appropriate

clusters. This algorithm finds the initial centroids by computing distances among each data object. Next it finds the closest pair of data objects add these in A1 data set and then delete them from the data object set D. This procedure is repeated until the number of objects in the set A1 reaches to threshold. At this point, generate another data-object set A2. Repeat this till 'k' such sets of data objects are obtained. At last, initial centroids are obtained by averaging all the vectors in each data-object set. This algorithm uses a heuristics to assign the data objects to cluster centroid.

Alternative KPSO-clustering (hybrid PSO and k-Means algorithm) method[19] to detect the cluster centroids of geometrical data sets automatic. This algorithm uses the special alternative metric to improve the traditional k-Means clustering algorithm to deal with various structure data sets.

The authors[20] have proposed a method issues criterion functions to select clusters to be split or merged to obtain a clustering structure that dynamically select cluster number. This approach can be highly effective to generate an initial clustering result with an automatically detected number of clusters as well as in incremental applications where the given cluster hierarchy should be updated dynamically as new documents are added or old documents are removed. Furthermore, it has been shown that online updates are favorable to batch updates, but compared with their increased computational power might not been applicable for real world data in the field of text mining without using the referenced heuristics. Concerning validity indices, the adapted Calinski Harabasz index and the simplified Bayesian Information Criterion lead to the best result to assess the clustering fitness.

III. Proposed Hierarchical K-Means Algorithm (HK-Means)

In this paper, we provide the implementation details of traditional k-Means algorithm with pre-processing task. Block diagram of this work is despite in Figure 4:

Figure 4: Block Diagram of proposed Hierarchical k-Means Algorithm

Platform used: VB 6.0 and MS SQL Server 5.0
Input: RIVER Dataset – n dimensional dataset
`

A. Transform Module: This module accept the RIVER dataset in text format and convert it in database file.

B. Pre-processing Module: Pre-processing is a very important step since it can affect the result of a clustering algorithm. This module calculates tuples with missing values and then provides options like maximum, minimum, constant or average for the treatment of missing values tuples before executing traditional k-Means algorithm. This process removes the noise in the dataset apply to k-Means algo.

C. Run Traditional k-Means Program: We implement the traditional k-Means algorithm that uses RIVER dataset which contains 133 tuples and 9 attributes. For traditional k-Means implementation we have consider RIVER dataset with 2 attributes and 133 tuples, number of clusters k = 2 and initial centroids are first two data objects with four iterations.

D. Propose Algorithm for Hierarchical k-Means(hk-Means): We propose an algorithm for hierarchical k-Means algorithm which will automatically detect initial centroids and preprocess dataset to remove noise represented in Figure 5.

Algorithm: Hierarchical k-Means **(hk-Means)**

Input: D is a Dataset
Output: k is number of clusters

Step 1: Read and Preprocess Dataset
Step 2: Calculate Mean M of the given D dataset
Step 3: Divide the D dataset into two Ranges: R1 and R2 data into R1 if they are lower than M and data into R2 if they are higher than M
Step 4: Apply k-Means Algorithm to each Ri (i=1,2)
Step 5: if results are satisfactory(based on MSE from step-4) go to step 2 else no further spliting possible and merge the previous clusters
Step 6: Do this process until no data object found in the list

Figure 5: Hierarchical k-Means(hk-Means) Algorithm

IV. Analysis and Result

If there are N tuples in the dataset, then, the similarity matrix can be computed in O(KNT). Let N is the number of tuples in the dataset. K is the number of clusters or centroids and T is the time to calculate the distance between to data objects. Time complexity of each iteration is O(KNT). There I number of iterations in k-Means algorithm. So, during I number of iteration the time complexity of this algorithm is O(IKNT). We have run traditional k-Means algorithm on RIVER dataset and the result is shown in Table 2:

TABLE -2 RESULT OF TRADITIONAL K-MEANS

Dataset	#Samples	#Iterations	#Clusters	MSE
RIVER	133	4	2	12.658
			3	5.622

Result of implemented traditional k-Means is presented in Figure-6.

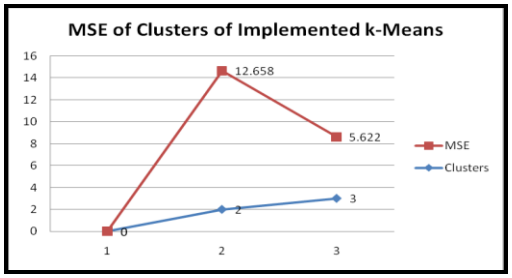

Figure 6: Result of Implemented k-Means Algorithm

V. Conclusion

Clustering is an important unsupervised technique in data mining to organize the data into groups having similarity. In clustering data are clustered based on certain criteria so data objects are closer to each other having similarity. Cluster analysis is the formal study of clustering, clustering algorithm and techniques. Clustering algorithm never is universal to solve all problems as they are designed on certain assumptions. Partitioned clustering algorithms are best suited for finding spherical shaped clusters in small to medium sized data. k-Means is the most popular clustering algorithm for its simplicity and favorable execution time. k-Means algorithm suffers from problems like to specify number of clusters and initial centroids in advance. Due to random initialization of initial centroids, k-Means always converges to local minima. Many researchers have worked and proposed many methods for automatic detection of initial centroids. We have analyzed various proposed methods for automatic detection of initial centroids in k-Means with pros and cons. Our propose method will enhance the k-Means for automatically detection of initial centroids with preprocessing task to remove noise. We have implemented traditional k-Means with various preprocessing tasks which remove noise in the implemented traditional k-Means algorithm.

VI. Future Work

Traditional k-Means clustering algorithm having problems of suppling number of clusters and initial centroids in advance. In this paper, we propose a method for automatic detection of initial clusters with the help of split and merge in traditional clustering algorithm. For this method, we have implemented traditional k-Means clustering algorithm with automatic pre-processing task to remove noise in the given dataset. We have apply this algorithm on RIVER dataset with number of clusters =2 and 3 and number of iterations = 4. In future, we will apply this algorithm on different datasets with changing number of clusters and iterations. We will also implement the modifications proposed in this paper and try to remove outliers presents in implmented k-Means algorithm.

References

[1] L.Kaufman and P. Rousseeuw," FindingGroups inData: An Introduction to Cluster Analysis", Wiley, 1990.

[2] T. Velmurugan and T. Santhanam," Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points ", Journal of Computer Science 6,Vol.3, pp. 363-368, 2010.

[3] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

[4] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006. Proceedings of the World Congress on Engineering 2009 Vol I.

[5] McQueen J, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., (1): pp. 281–297, 1967.

[6] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases .

[7] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.

[8] Vaishali R. Patel, Rupa G. Mehta,"Clustering Algorithms: A Comprehensive Survey", International Conference on Electronics, Information and Communication Systems Engineering (ICEICE 2010), March 28th – 30th, 2011, Jodhpur, Rajasthan.

[9] A. Jain, R. Duin, and J.Mao, "Statistical Pattern Recognition: A Review",IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37, 2000.

[10] J. Kleinberg, "An impossibility theorem for clustering," in Proc. 2002, Conf. Advances in Neural Information Processing Systems, vol. 15, pp. 463–470, 2002.

[11] Chris Ding and Xiaofeng He, "Cluster merging and splitting in hierarchical clustering algorithms", IEEE International Conference on Data Mining (ICDM'02),2002.

[12] L.Kaufman and P. Rousseeuw," FindingGroups inData: An Introduction to Cluster Analysis", Wiley, 1990.

[13] K. A. Abdul Nazeer, M. P. Sebastian," Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering 2009 Vol I,WCE 2009, July 1 - 3,2009,London,UK

[14] Tajunisha N, Saravanan V," A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets",International Journal of Advanced Science and Technology,Vol 27, February,2011

[15] Jain, A., Dubes, R.,: Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ , 1988.

[16] Tajunisha, Saravanan,"Performance analysis of k-means with different initialization methods for high dimensional data", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010

[17] Fahim A.M,Salem A.M, Torkey F. A., Saake G and Ramadan M.A: An Efficient k-means with good initial starting points, Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol.2, No. 19, pp. 47-57, 2009

[18] Madhu Yedla et al. (2010) : "Enhancing K-means clustering algorithm with improved initial centers",International Journal of Computer Science and information Technologies. Vol.1(2), pp. 121-125, 2010

[19] Fun Ye,Ching-Yi Chen,"Alternative KSPO-clustering Algorithm", Tamkang Journal Of Science and Engineering, Vol. 8, No.2, pp. 165-174, 2005

[20] Markus Muhr, Micheal Granitzer," Automatic Cluster Number Selection using a Split and Merge K-Means Approach", DEXA Workshop,, ISBN: 978-0-7695-3764-4, pp 363-367, 2009

[21] Zhong Wei, et al. "Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property." *IEEE Transactions on Nanobioscience*. Vol. 4. No. 3., pp. 255-265, 2005