Enhancement in K-Mean for Data clustering: A Review

Preeti Puri1, Isha Sharma2
1Post Graduate Student, Dept. of Computer Science Engineering (Software Engineering), Chandigarh University, Mohali, Punjab, 140413
Er.preeti1991@gmail.com
2Assistant Professor, Dept. of Computer Science Engineering, Chandigarh University, Mohali, Punjab, 140413
Ishasharma211@gmail.com

ABSTRACT-Clustering is the most commonly used method for grouping of related observations in a data set. The K-Means method is one of the mostly used clustering techniques for a variety of applications like defect detection, networking etc. In this paper we are proposing a method for making the K-Means algorithm more efficient and effective so as to reduce the complexity. Clustering algorithm forms a vector of topics for each document and measures the weights of how healthy the document fits into each cluster. Clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. Clustering is an unsupervised classification method aiming at creating groups of objects, or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. In this paper the Cluster analysis or clustering is used so as to set the objects into group clusters and then those clusters into sub clusters.

Index Term: - Data mining, Requirement analysis, clustering, k-mean.

I. INTRODUCTION:

Requirement analysis:

Software engineering task helps us to fill the gap between system requirements engineering and software design. It provides software designer with a model of system information, function and behaviour. In this it expect to do a little bit of design during analysis and a little bit of analysis during design. The main Objectives of software analysis is to identify customer's needs, evaluate system for feasibility, perform economic and technical analysis, allocate functions to system elements, establish schedule and constraints, and create system definitions.
Data mining uses the automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse.

Clustering is the method to cluster the data into the sub clusters according to their mean values. The goal of clustering is descriptive to discover a new set of categories. Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between them and dissimilar compared to objects of other groups. Cluster analysis is a very important technology in Data Mining. It divides the datasets into several meaningful clusters to reflect the data sets natural structure. Cluster is aggregation of data objects with common characteristics based on the measurement of some kind of information.

INITIAL CLUSTER:

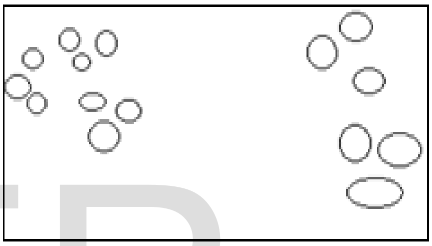

Fig: 1 Initial cluster in the data warehouse

In above figure the data is in the normal form with no proper division. This data is in the simpler form in the data warehouse.

TWO CLUSTERS:

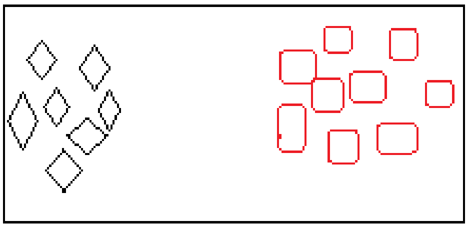

Fig: 2 Data is divided into the two clusters

In this figure the data is divided into the clusters. The data is clustered on the basis of the similar data points.

FINAL CLUSTER:

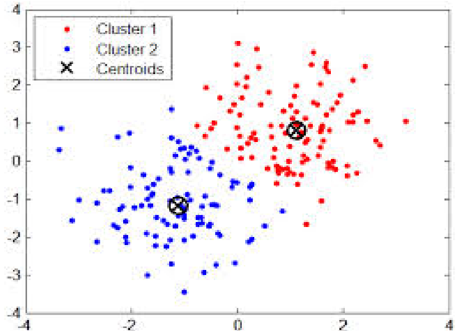

Fig: 3 Clustered data with mid points

In above Fig 3 we have clustered the data by finding out the data values and on the basis of that clustering is done. With clustering we can easily differentiate between the data and can also find the centroids.

II. LITERATURE SURVEY

Clustering technique is mainly grouping the data into modules and then those modules into the nearest mean by finding out the mean values of the modules.

[1]Comparison the various clustering algorithms of weka tool

Narendra Sharma, Aman Bajpai. In this paper they are working on various clustering algorithms. Their main aim is to show the comparison of the different clustering algorithms on WEKA and check out that which method is the more efficient method.

[2]Performance analysis of k-means with different initialization methods for high dimensional data

Tajunisha and Saravanan: In this paper, they have analyzed the performance of their proposed method with the existing works. In this they have used Principal Component Analysis (PCA) for dimension reduction and to find the initial centroid for k-means and then they have used heuristics approach to reduce the number of distance calculation to assign the data point to cluster. By comparing the results on iris data set, it was found that the results obtained by the proposed method are more effective than the existing method.

[3]A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set

D.Napoleon, S.Pavalakodi. K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply PCA on original data set and obtain a reduced dataset containing possibly uncorrelated variables. In this paper they have used principal component analysis and linear transformation for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm.

[4]Evolving limitations in K-means algorithm in data mining and their removal

” Kehar Singh, Dimple Malik and Naveen Sharma: K-means is very popular because it is conceptually simple and is computationally fast and memory efficient but there are various types of limitations in k means algorithm that makes extraction somewhat difficult. In this paper they have discussed the limitations and how these limitations will be removed.

[5]A Modified K-Means Algorithm for Circular Invariant Clustering

Dimitrios Charalampidis Member: This paper introduces a distance measure and a K-means-based algorithm, namely, Circular K-means (CK-means) to cluster vectors containing directional information, such as Fd, in a circular-shift invariant manner. A circular shift of Fd corresponds to pattern rotation, thus, the algorithm is rotation invariant. An efficient Fourier domain representation of the proposed measure is presented to reduce computational complexity. A split and merge approach (SMCK-means), suited to the proposed CK-means technique, and is proposed to reduce the possibility of converging at local minima and to estimate the correct number of clusters. Experiments performed for textural images illustrate the superior performance of the proposed algorithm for clustering directional vectors Fd, compared to the alternative approach that uses the original K-means and rotation-invariant feature vectors transformed from Fd.

[6]A Modified k-means Algorithm to Avoid Empty Clusters

Malay K. Pakhira: This paper presents a modified version of the k-means algorithm that efficiently eliminates this empty cluster problem. They have shown that the proposed algorithm is semantically equivalent to the original k-means and there is no performance degradation due to incorporated modification. Results of simulation experiments using several data sets prove our claim.

[7]An Efficient k-Means Clustering Algorithm: Analysis and Implementation

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y.Wu: In this paper, they presented a simple and efficient implementation of Lloyd's k means clustering algorithm, which they call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. They established the practical efficiency of the filtering algorithm in two ways. First, they presented a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, they presented a number of empirical studies both on synthetically generated data and on real data sets from applications in colour quantization, data compression and image segmentation.

[8]Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation

Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa: This paper presents an improved version of the Moving K Means algorithm called Enhanced Moving K-Means (EMKM) algorithm. In the proposed EMKM, the moving concept of the conventional Moving K-Means is enhanced. Two versions of EMKM, namely EMKM-1and EMKM-2 are proposed. The qualitative and quantitative analyses have been performed to measure the efficiency of both EMKM algorithms over the conventional algorithms (i.e. K-Means, Moving K-Means and Fuzzy C-Means) and the latest clustering algorithms (i.e. AMKM and AFMKM). They investigated that the proposed algorithms significantly outperform the other conventional clustering algorithms.

III. CLUSTERING METHODS:

Han and Kamber (2001) cluster the methods into three categories: density based, model based and grid based method. Clustering methods also had an alternative categorization based on induction principle which is presented in (Estivill-Castro, 2000).

1. Hierarchical Methods:

In these techniques two approaches are used top down approach and bottom up approach. These can be sub divided into:

a. Agglomerative hierarchical: In this technique each object in the cluster represents the cluster of its own. Then these clusters are merged to get a desired cluster.

b. Divisive hierarchical clustering: Initially all the objects belongs to the same cluster. Then the cluster is subdivided into the sub clusters and then again these sub clusters into their sub clusters. This process is repeated until the desired structure is obtained. The hierarchical is divided according to the manner until the similarity measure is calculated (Jain et al., 1999):

Single-link clustering (also known as minimum method or correctness or nearest neighbor method):In this they consider the distance between the two clusters which is equal to the shortest distance from the one object of one cluster to the other object of the other cluster . In this if the data contain the similar clusters then the similarity of one cluster is equal to similarity from any member of one cluster to the any member of another cluster (Sneath and Sokal, 1973).

Complete-link clustering (also known as the maximum method or diameter or further neighbor method):In this method we consider the distance between the two clusters which is to be equal to the longest distance of any cluster of any member of one cluster to the other member of another cluster (King,1967).

Average-link clustering (also known as minimum variance method) –In average link clustering method they consider the distance between the two clusters which is to equal to the average distance between from the member of one cluster to the member of the cluster. Such clustering algorithms may be found in (Ward, 1963) and (Murtagh, 1984).

2. Partitioning Methods:

Partitioning methods is the method to relocate the instances by moving them from one cluster to the another cluster by starting from the root partitioning .In partitioning method, the number of clusters that are pre-set by the user is required. In this method to achieve the global optimality we use the following methods:

a. Error Minimization Algorithms.

These are the algorithms that tend to work in all the clusters whether it is isolated or compact clusters. The main aim in this is to minimize the error criterion which is responsible for measuring the distance of each the instance to its representative value. The well known criterion is Sum of Squared Error (SSE) that measures the total squared Euclidian distance of instances to its representative values. The most commonly used algorithm which deals with the squared distance is the K-Mean algorithm. This algorithm partition the data into the k clusters say (C1, C2……C_K) Represented by the centre of their means. [9]

The centre of cluster is calculated by calculating the mean of all the instances that belong to the cluster.

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} xq$$

Here N_k is the number of instances belonging to the cluster and ʹk is the mean of the cluster. [9]

Input: S (instance set), K (number of cluster)

Output: clusters

```
1: Initialize K cluster centres.
2: while termination condition is not satisfied do
3: Assign instances to the closest cluster centre.
4: Update cluster centres based on the assignment.
5: end while
```

The gradient decent procedure can be viewed in the K-mean algorithm. It begins with the set of k cluster centre and then updates itself to decrease the error function in the clusters. [9]

3. Density-based Methods:

Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution (Banfield and Raftery, 1993). The overall distribution of the data is assumed to be a mixture of several distributions. The aim of these methods is to identify the clusters and their distribution parameters.

4. Model-based Clustering Methods:

This method helps in optimizing the fit between the models and the given data. Unlike in conventional clustering it identifies the group of objects, model based clustering and the find out the characteristics of each group where it represent as concept or class.

5. Grid-based Methods:

These methods partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time (Han and Kamber, 2001).

6. Soft-computing Methods:

The usage of neural networks in clustering tasks. This section discusses the usefulness of other soft-computing methods in clustering tasks.

It provides rapid dissemination of important results in soft computing technologies, a fusion of research in evolutionary algorithms and genetic programming, neural science and neural net systems, fuzzy set theory and fuzzy systems, and chaos theory and chaotic systems. The linkage of ideas and the techniques of this method with other, it serves as the platform for the extensions and new applications. A journal is the forum for all the engineers and scientists to be engaged in development and research in the fast growing field. [10]

IV. RELATED WORK:

Software engineering is all about the production of software, from its initial stage to its final stage. Computer science is related to the software engineering in many aspects. As computer science is used to provide the discipline that uses for the theory basis professions in software engineering. The nature and the complexity of the software system are changing. In the today's world the applications that are used to develop the software system are more complex, because in these applications GUI i.e. graphical user interface and client server architecture is used. Software are distributed among the various system, hence it can works in one or more than one processes. A software engineering is related to all the aspects that are used in the software production. Software is basically a generic term, which is used for organizing the data and instructions that are collected to develop it. The software is broken into the two categories: system software and application software. The system software is used to manage the hardware components, so that other software or user sees it as a functional unit. The software contains the operating system and some more utilities like disk formatting, file managers, display managers, etc. A software component model specifies standards for composition of and interaction between software components. To facilitate the use of such models, dedicated software tools and infrastructures are often implemented. These may include run-time environments for component execution and interaction as well as tools for component development, composition, and deployment. A software component technology is a set of dedicated software products supporting the use of a specific software component model. Heineman and Council use the term component model implementation to denote the run-time parts of a software component technology. The large software's have large and complex structures due to which it have many modules. To properly classify the categories of the modules various techniques had been proposed in the previous times among all the proposed techniques clustering is the most efficient technique for clustering the similar type of functions. In the base paper, genetic algorithms had been applied for clustering the similar type of data. The genetic algorithms are based on the chromosome values, which is the inefficient technique of clustering. In this work, we will propose another clustering technique for clustering of similar type of functions.

V. PROPOSED WORK:

In our work we will use Mat lab tool to perfectly differentiating the clusters. So that we can know about the nearest clusters which are lying closer.
In this
1. the data
2. Analyze the data
3. Apply supervised and unsupervised learning
4. Apply the k-mean clustering and record the results
5. Implement the existing and new algorithm to differentiate between the data easily.

Step I: Identify the data

First step is to identify the data. In this we will check the data that what are the main constituents of the data which are dependent and which are independent.
a. Check the data.
b. Identify the factors on which data is dependent.

Step II: Analyze the data.

Second step is to analyze the data properly so that no factor should be left.

Step III: Apply supervised learning and unsupervised learning [11].

Supervised learning is the learning in which the data is extracted from the target class. The methods used in this are:

a. Decision trees:

Decision Trees are helpful in deciding the path which is easier. They provide a highly effective structure with which we can check various options and investigate the possible outcomes of choosing those options. Decision tree provide a balanced picture of the risks and rewards with each action performed.

b. Random Forest:

In this they assumed that the user knows about the construction of single classification trees. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest [12]. Each tree gives a classification about the data.

c. Naive bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable and a dependent feature vector through, Bayes' theorem states the following relationship: [13]

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}$$

In **unsupervised learning** there is no previous information is there. Everything is done dynamically. In this the main methods are:

a. K-mean clustering:

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as BabyCenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop we may notice that the k centres change their location step by step until no more changes are done or in other words centres do not move any more[14]

$$J\,(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

b.Hierarchical clustering:

Hierarchical clustering group's data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows you to decide the level or scale of clustering that is most appropriate for your application. It incorporates the pdist, linkage, and cluster functions, which you can use separately for more detailed analysis. The dendrogram function plots the cluster tree. [15]

c. Density based algorithm:

Clustering based on density (local cluster criterion), such as density-connected points or based on an explicitly constructed density function [16]

Major features of this technique are to discover clusters of arbitrary shape and to handle noise. In this we need the density parameters and one scan technique.

Step IV: Apply the k-mean clustering

Fourth step is to apply the k-mean clustering:

In this technique we will partition the data into 'n' observations and then partition the data into 'k' clusters. After that the data is grouped in minimizing the sum of squares of distance between the data and finding the center point of the cluster.

So that the division can be done easily.

Step V: Implement the existing and new algorithm to differentiate between the data easily.

Final step is to implement the algorithm which is existing and the new algorithm on which we have proposed our work. With the existing and new we can differentiate the data easily and find out that how efficient the results will be.

VI. CONCLUSION

The k-mean algorithm follows a simple way to classify a given data set through a certain number of clusters. From this we have concluded that k-mean is the most efficient method. This method is used in the daily life and the results with this algorithm are more superior with comparison to other algorithms. This algorithm has high accuracy and optimal results are produced in this.

VII. FUTURE WORK:

In our work we hope that researchers do certain experiments in this field and discover some superior methods which results more efficient than this. There are many learning methods which can produce better results. Data mining is a dynamic field on which the data miners are working and this field is continuously developed. Data flow diagrams in this can help in easy differentiation of the data.

REFERENCES:

[1] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya," Comparison the various clustering algorithms of weka tools"," International Journal of Emerging Technology and Advanced Engineering" (ISSN 2250-2459, Volume 2, Issue 5, May 2012)

[2]Tajunisha and Saravanan, "Performance analysis of k-means with different initialization methods for high dimensional datasets, "International Journal of Artificial Intelligence & Applications (IJAIA), vol. 1, no.4, pp.44-52, Oct. 2010.

[3] D.Napoleon, S.Pavalakodi,"A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set," International Journal of Computer Applications (0975– 8887), vol. 13, no.7, pp.41-46, Jan 2011.

[4] Kehar Singh, Dimple Malik and Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal,"IJCEM International Journal of Computational Engineering &Management, vol. 12, pp.105-109, Apr. 2011.

[5] Dimitrios CharalampidisI,"A Modified K-Means Algorithm for Circular Invariant Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp.1856-1865, Dec 2005.

[6] Malay K. Pakhira,"A Modified k-means Algorithm to Avoid Empty Clusters," International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp.220-226, May 2009.

[7] TapasKanungo, David M.Mount, Nathans. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y.Wu,"An Efficient-Mean Clustering Algorithm: Analysis an Implementation,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891, Jul 2002.

[8] Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa, "Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation,"IEEE, pp.833-841.

[9] Data mining and knowledge discovery handbook, Lior Rokach (Department of Industrial Engineering) Tel-Aviv University-Chapter15, CLUSTERING METHODS

[10] Soft Computing, A Fusion of Foundations, Methodologies and Applications

[11] Amerada Chug1, Shafali Dhall,"Software Defect Prediction Using Supervised Learning Algorithm and Unsupervised Learning Algorithm" (USICT, GGSIPU, New Delhi)

[12]https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

[13]http://scikit-learn.org/stable/modules/naive_bayes.html

[14]https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[15]http://in.mathworks.com/help/stats/hierarchical-clustering.html

[16] www2.cs.uh.edu/~ceick/ML/Topic9

[17]N.S.Chandolikar, V.D.Nandavadekar,"Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset, "International Journal of Computer Science and Engineering(IJCSE),vol.1,pp.81-88,Aug 2012.

[18] Mehmet Koyuturk, Ananth Grama and Naren Ramakrishnan,"Compression, Clustering and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets,"IEEE Transactions on Knowledge and A Data Engineering", vol. 17, no. 4, pp.447-461, Apr 2005.

[19] Christopher M. Bishop and Michael E. Tipping, "A Hierarchical Latent Variable Model For Data Visualization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp.281-293, Mar. 1998.

[20]Mu-Chun Su and Chien-Hsing Chou, "A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.