Current Trends in Technology and Science
ISSN: 2279-0535, Volume: VI Issue: II

Enhanced Clustering Based on K-means Clustering Algorithm and Proposed Genetic Algorithm with K-means Clustering

Pradeep Salve
M. Tech. Scholar
Department of IT
U.I.T., Barkatullah University
Bhopal, M.P., India
salvepradeepnri@gmail.com

Dr. Poonam Sinha
Head of Department
Department of EC
U.I.T., Barkatullah University,
Bhopal, M.P., India
poonamuit@yahoo.com

Prof. Rachna Kulhare
Assist. Professor,
Department of IT
U.I.T., Barkatullah University,
Bhopal, M.P., India
rachna12kulhare@gmail.com

Abstract--In this paper targeted a variety of techniques, tactics and distinctive areas of the studies that are useful and marked because the crucial discipline of information mining technologies. The overall purpose of the system of statistics mining is to extract beneficial facts from a large set of information and changing it right into a shape that is comprehensible for in addition use. Clustering is an important chore in information evaluation and data mining applications. Statistics divides into similar item businesses based on their functions by clustering method. Every records organization with similar objects is clusters. Clustering algorithms have many classes like hierarchical, partition, density-primarily based and grid based totally. Partition-based clustering is centroid based which splits information factors into k partition and each partition represents a cluster. K-means is a clustering set of rules that is used widely. on this paper, do a evaluation on k-method clustering on this paper numerous changed K-means algorithm are discussed which take away the difficulty of purpose algorithm improve the rate and efficiency.

Keywords: - Data Mining, Clustering**,** Clustering, Genetic algorithm, K-means Algorithm, Partition-based clustering.

I. INTRODUCTION

Clustering is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, but data belonging to different cluster differ. A cluster is collections of data object that are similar to one another are in same cluster and dissimilar to the objects are in other clusters. The demand for organizing the sharp increasing data and learning valuable information from data, which makes Clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, and statistics and so on. Cluster analysis is a tool that is used to observe the characteristics of cluster and to focus on a particular cluster for further analysis. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, Analysis the data by application software, Present the data in a useful format, such as a graph or table[1]**.**

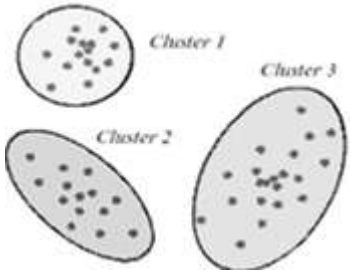

Figure1 Clustering

II. DIFFERENT TYPES OF CLUSTERS

1. Contiguous clusters: Nearest neighbor or transitive a cluster is a group of points such that a single point in a cluster is closer to one or more other points in the cluster than to any other point not in the cluster.

2. Centre based clusters: A cluster is a group of objects so that an object in a cluster is closer to the centre of a cluster, than to the centre of other cluster. The centre of a cluster is called a centroid, the average of all the points in the cluster, or a medoid, the most representative point of a cluster

3. Density- based clusters: A cluster is dense region of points, which is individual separated by low-density regions, from the other regions of high density regions. It used when the clusters are very irregular, and when noise and outliers are available

4. Well-separated clusters: A cluster is a collection of points such that any other point in a cluster closer or more similar to each and every other point in the cluster than to any point not in the cluster [2].

Clustering: Cluster analysis is explored the structure of data. Core Cluster analysis is a clustering. Clustering analysis in a data is an unknown label class .So it is learned by observation not learned by example. Clustering divide the data set into classes using the principle of "Maximum intra class similarity and Minimum inter class similarity". It doesn't have any assumption about the category of data. The basic clustering techniques are Hierarchical, Partitioning, Density based, Grid based and Model based clustering. Some sort of measure that can determine whether two objects are similar or dissimilar is required to add them into particular class. The distance measuring type varies for different attribute type. Clustering can also used to detect outline in data which may occur due to human error or some abnormal events occurred while creating data set .Cluster work well on scalable,

heterogeneous and high dimensional data set. In all the clustering algorithms user defined parameters are given as input to find either similarity, dissimilarity among clusters and for root attribute of cluster and for maximum or minimum number of clusters[3].

Partition-based clustering: It is centroid based clustering in which data pointsspilits into k partition and each partition represents a cluster. Different methods of partitioning clustering are k-means, bisecting k-means method, Medoids method, Portioning around Medoids (PAM), CLARA (Clustering LARge Applications) and the Probabilistic centroid. K-means clustering: K-means clustering technique is a technique of clustering which is widely used. This algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition end observations into k cluster. Which each observation belongs to the cluster with the nearest mean. The basics algorithm is very simple 1. Select K points as initial centroids. 2. Repeat 3. Form K cluster by assigning each point to its closest centroid. 4. Recomputed the centroid of each cluster until centroid does not change. Properties of k-means algorithms are 1. Large data set are efficiently processed. 2. It often terminates at a local optimum. 3. It works only on numeric values. 4. The shape of clusters is convex [4].

III. LITERATURE SURVEY

Hareesha et al. [5] present a modified K-means algorithm to improve the cluster quality and to fix the optimal number of cluster. As input number of clusters (K) given to the K-means algorithm by the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The method proposed in this paper works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. The new cluster centers are computed by the algorithm by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. This algorithm will overcome this problem by finding the optimal number of clusters on the run. The proposed approach takes more computational time than the K-means for larger data sets. It is the major drawback of this approach.

Libao ZHANGet al. [6]propose a simple and qualitative methodology using k means clustering algorithm to classify NBA guards and used the Euclidean distance as a measure of similarity distance. This work used k-Means clustering algorithm and 120 NBA guards' data. Manual classification of traditional methods is improved using this model. According to the existing statistical data, the NBA players are classified to make the classification and evaluation objectively and scientifically. This work shows that this is very effective and reasonable methodology. Therefore, based on classification result the guards' type can be defined properly. Meanwhile, the guards' function in the team can be evaluated in a fair and objective manner.

J. K. Sahiwal et al. [7]. Enhanced the traditional k-means by introducing Ranking method. Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means. The Ranking Method is a way to find the occurrence of similar data and to improve search effectiveness. The tool used to implement the improved algorithm is Visual Studio 2008 using C#. The advantages of k-means are also analysed in this paper. The author finds k-means as fast, robust and easy understandable algorithm. He also discuss that the clusters are non-heirarchical in nature and are not overlapping in nature. The process used in the algorithm takes student marks as data set and then initial centroid is selected. Euclidean distance is then calculated from centroid for each data object. Then the threshold value is set for each data set. Ranking Method is applied next and finally the clusters are created based on minimum distance between the data point and the centroid. The future scope of this paper is use of Query Redirection can be used to cluster huge amount of data from various databases.

Philip K. Maini et al. [8] described a useful survey of fuzzy clustering in main three categories. The first category is basically the fuzzy clustering depends on exact fuzzy relation. The second one is the fuzzy clustering based on single objective function. Finally, it is given an overview of a nonparametric classifier. That is the fuzzy generalized k nearest neighbor rule. The fuzzy clustering algorithms have obtained great success in a variety of substantive areas

Md. M. Rahman et al. [9] gave an algorithm to compute better initial centroids based on heuristic method. The newly presented algorithm results in highly accurate clusters with decrease in computational time. In this algorithm author firstly compute the average score of each data points that consists of multiple attributes and weight factor. Sorting technique is applied to sort the output that was previously generated. The data points are then divided into k cluster i.e. number of desired cluster. Finally the nearest possible data point of the mean is taken as initial centroid. Experimental outputs show that the algorithm reduces the number of iterations to assign data into a cluster. But the algorithm still deals with the problem of assigning number of desired cluster as input.

J. Wang et al. [10] discuss an improved k-means clustering algorithm to deal with the problem of oulier detection of existing k-means algorithm. The proposed algorithm uses noise data filter to deal with this problem. Density based outlier detection method is applied on the data to be clustered so as to remove the outliers. The motive of this method is that the outliers may not be engaged in computation of initial cluster centres. In the next step fast global k-means algorithm proposed by Aristidis Likas is applied to the output generated

previously. The results between k-means and improved k-means are compared using Iris, Wine, and Abolone datasets. The Factors used to test are clustering accuracy and clustering time. The disadvantage of the improved k-means is that while dealing with large data sets, it will cost more time

S. Jasola et al. [11] proposed a new improved algorithm named as Boundary Restricted Adaptive Particle Swam Optimization (BRAPSO) algorithm with boundary restriction strategy for particles that travel outside the boundary search space during PSO process. Nine data sets were used for the experimental testing of BR-APSO algorithm, and its results were compared with PSO as well as some other PSO variants namely, K-PSO, NM-PSO, and K-Means clustering algorithms. It has been found that the proposed algorithm is robust, generates more accurate results and its convergence speed is also fast as compared to other algorithms.

J. Fan et al. [12] proposed a particle swarm optimization approach with dynamic neighborhood based on kernel fuzzy clustering and variable trust region methods (called FT-DNPSO) for large-scale optimization. It adaptively adjusts the initial region and clusters different dimension into groups, which expedites convergence and search in the effective range. The adaptive strategy avoids or alleviates the prematurity of the PSO algorithm. The simulation results, with eight classical benchmark functions, twenty CEC2010 test ones and soft computing special session test; demonstrate that the proposed FT-DNPSO outperformed other PSO algorithms for large-scale optimization.

Wen-Jun Zhang et al. [13] worked out an adaptive particle swarm optimization (PSO) on individual level. By analyzing the social model of PSO, a replacing criterion based on the diversity of fitness between current particle and the best historical experience is introduced to maintain the social attribution of swarm adaptively by removing inactive particles. Three benchmark functions were tested which indicates its improvement in the average performance.

Garima Mishra et al. [14] proposed a Linear PCA based hybrid K-Means clustering and PSO algorithm (PCA-K-PSO). In (PCA-K-PSO) algorithm the fast convergence of K-Means algorithm and the global searching ability of Particle Swarm Optimization (PSO) are combined for clustering large data sets using Linear PCA. Better clustering results can be obtained with PCA-K-PSO as compared to ordinary PSO. This was effectively developed in order to make its use for efficient clustering of high-dimensional data sets.

Britos et al. [15]. the image compression problem using genetic clustering algorithms based on the pixels of the image was proposed in. GA was used to obtain an ordered representation of the image and then the clustering was performed to obtain the compression.

Liu Xumin et al. [16] present the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in iteration. This repetitive process affects the efficiency of clustering algorithm. An improved k-means algorithm is proposed in this paper. A simple data structure is required to store some information in iteration which is to be used in the next iteration. Computation of distance in iteration is avoided by the proposed method and saves the running time. The work of paper shows that proposed method can effectively improve the speed and accuracy of clustering, reducing the computational complexity of the k- means.

IV. RESULT ANALYSIS

Data mining and identify various challenges in the field data mining. Error free data in clustering by definition and Minimum cluster and useful information an increase accuracy clustering technique. A efficiently cluster according to their importance. MATLAB is the high level language and interactive environment used by millions of engineers and scientists worldwide. It lets explore and visualize ideas and collaborate across different disciplines with signal and image processing, communication and computation of results. MATLAB provides tools to acquire, analyze, and visualize data, enable you to get insight into your data in a division of the time it would take using spreadsheets or traditional programming languages. It can also document and share the results through plots and reports or as published MATLAB code. MATLAB (matrix laboratory) is a multi paradigm numerical computing situation and 4th generation programming language. It is developed by math work; MATLAB allows matrix strategy, plotting of function and data, implementation of algorithm, construction of user interfaces with programs. MATLAB is intended mainly for mathematical computing; an optional tool box uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities. It is simulating on mat lab 7.8.0 and for this work we use Intel 1.4 GHz Machine. MATLAB is a high-level technical compute language and interactive environment for algorithm development, data visualization, records analysis, and numeric computation Mat lab is a software program that allows you to do data manipulation and visualization, calculations, math and programming. It can be used to do very simple as well as very sophisticated tasks are Database, analysis, visualization, and algorithm development. You can perform efficient data retrieve enhancement. Many functions in the toolbox are multithreaded to take benefit of multicore and multiprocessor computers. An additional package, Simulink, natural network is the key features of MATLAB, a high level language. It is processes used Dataset. Based result graph show below .K-mean algorithm is more error in dataset clustering as compare to PA. Is better as compare to k-means because k-mean dataset cluster error rate is

more but is minim dataset cluster error. K-mean algorithm is time take minim as compare to PA. Dataset based result graph show below. K-mean algorithm is more error in dataset clustering as compare to PA better as compare to k-means because k-mean dataset cluster error rate is more but is minim dataset cluster error. K-mean algorithm is time take minim as compare to .The graph shows above with the purpose algorithm high accuracy as compared to the normal pervious algorithm K-Mean has because the data set has less error in dataset also called filtered data set.k-means values 4.53561% and PA values 1.50158 % minimum. It is good proposed algorithm.

V. CONCLUSION

Data mining using clustering in k-means is still at the stage of searching and improvement. K-means clustering techniques of data mining are analyzed. This work shows that there are several methods to improve the clustering with different approaches. Various clustering techniques are reviewed which improve the existing algorithm with different Perspective The result concludes that many improvements are basically required on k-means to improve problem of cluster initialization, cluster quality and efficiency of algorithm. On the other hand in this paper a study the three clustering algorithms one is, simple K-Means partitioning algorithm**,** the hybrid algorithm which is the combination of simple K-Means. K-means is with PGA to get the optimize no. of clusters from the result of simple K Means algorithm .Both algorithm are simple to understand and can be applicable for various type of data like medical data set, numerical data set. K-means values 4.53561% and PA values 1.50158 % minimum. It is good proposed algorithm. Some limitations of existing algorithm will be eliminated in future work. This technique will be useful in extraction of useful information using cluster from huge Database.

REFERENCES

[1]. Malwindersingh, Meenakshibansal,"A Survey on Various K- Means algorithms for Clustering", IJCSNS International Journal of Computer Science and Network Security, VOL.15 No.6, June 2015.

[2]. Shaheda Akthar, Sk.Md.Rafi, "Improving the Software Architecture through Fuzzy Clustering Technique", Vol 1 No 154-57, 2011.

[3]. Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", India, MK Publications, 2006.

[4]. Amandeep Kaur Mann, Navneet Kaur Mann, "Review Paper On Clustering Techniques" ,Global Journal Of Computer Science And Technology Software & Data Engineering, VOL. 13 ,2013.

[5]. Shafeeq, A., Hareesha, K., "Dynamic Clustering of Data with Modified K-Means Algorithm", International Conference on Information and Computer Networks, vol. 27, 2012.

[6]. Libao ZHANG, Faming LU, An LIU, Pingping GUO, Cong LIU, "Application of K-Means Clustering Algorithm for Classification of NBA Guards", International Journal of Science and Engineering Applications Volume 5 Issue 1, ISSN- 2319-7560 (Online), 2016.

[7]. Jaspreet Kaur Sahiwal, Navjot Kaur, Navneet Kaur "Efficient Kmeansclustering Algorithm Using Ranking Method In Data Mining", ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[8]. Don Kulasiri, Sijia Liu, Philip K. Maini and RadekErban, "Diffuzzy: A fuzzy clustering algorithm for complex data sets", International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol.1, no.4,pp. 402-417, 2010.

[9]. Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md. Nasim Akhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", 2012 7th International Conference on Electrical and Computer Engineering 20-22, Dhaka, Bangladesh, IEEE, December, 2012.

[10]. Juntao Wang & Xiaolong Su "An improved K-Means clustering algorithm, IEEE, 2011.

[11]. S. Rana, S. Jasola, and R. Kumar, "A boundary restricted adaptive particle swarm optimization for data clustering," International Journal of Machine Learning & Cyber. Springer, pp.391-400, June 2012.

[12]. Jianchao Fan, Jun Wang, and Min Han, "Cooperative Convolution for Large-scale Optimization Based on Kernel Fuzzy Clustering and Variable Trust Region Methods," IEEE Transactions on TFS-2013-0157, pp. 1-12, 2013.

[13]. Xiao-Feng Xie, Wen-Jun Zhang, and Zhi-Lian Yang, "Adaptive Particle Swarm Optimization on Individual Level," IEEE, International Conference on Signal Processing (ICSP), Beijing, China, 2002, pp. 1215-1218.

[14]. Chetna Sethi and Garima Mishra, "A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset," International Journal of Scientific & Engineering Research, Volume 4, Issue 6, pp.1559-1566, June-2013.

[15]. Merlo, Caram, Fernández, Britos, Rossi, &García Martínez.," Based Image compression Genetic Algorithm", SBAISimpósio Brasileiro de Automa Inteligente, São Paulo, SP, 08-10 de Setembro de 1999.

[16]. Shi Na, Liu Xumin, Guan Yong, "Research on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm", Intelligent Information Technology and Security Informatics,2010 IEEE Third International Symposium on, (pp. 63-67),2-4 April, 2010.