International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 3, Issue 10, October 2016)

A Survey on Identifying Cluster Center for Different Dataset Using K-means Clustering Algorithm

Abhilasha Patel[1]
M. Tech. Scholar Department of CSE
TIT, R.G.P.V., Bhopal
M.P., India
abhilasha8874patel@gmail.com

Prof. Kamlesh Chandravanshi[2]
Department of CSE
TIT, R.G.P.V., Bhopal
M.P., India
kamlesh.vjti@gmail.com

Prof. Deepak Tomar[3]
Department of CSE
TIT, R.G.P.V., Bhopal
M.P., India
tomar_deepak01@yhoo.in

***Abstract*--Data mining is used clustering algorithm based on the K-means clustering and improvement center in cluster may be done by selecting acceptable initial cluster centers to converge quickly to the local optimum. Within the proposed work clustering algorithm this paper deeply works over the fact that the k-means c1usteriug algorithm is extremely sensitive to the initial values but new propose clustering algorithm is best. So as to enhance the dependence on the initial values, it proposes a clustering algorithm supported enhanced center. Through continuous modification to density threshold, it gets the more clustering centers, and merges them till the specified number of clustering center and algorithm is applied to the various dataset for clustering analysis, so the result proves that the improved algorithm optimizes the dependence; Finally, achievement best knowledge in dataset it's also analysis of the most effective centroids. In clustering best data acaccuracy of the proposed algorithm is best. In clustering best data and minimize cluster error through proposed algorithm.**

Keywords-Clustering, Data Mining, Clustering, Genetic algorithm-K-Means (GAKM), Centroid, K-means Algorithm,

I. INTRODUCTION

Clustering techniques have become very popular in a number of areas, such as engineering, medicine, biology and data mining [1]. A good survey on clustering algorithms can be found in [2]. The k-means algorithm [3] is one of the most widely used clustering algorithms. The algorithm partitions the data points (objects) into C groups (clusters), so as to minimize the sum of the (squared) distances between the data points and the center (mean) of the clusters. In spite of its simplicity, the k-means algorithm involves a very large number of nearest neighbor queries. The high time complexity of the k-means algorithm makes it impractical for use in the case of having a large number of points in the data set. Reducing the large number of nearest neighbor queries in the algorithm can accelerate it. In addition, the number of distance calculations increases exponentially with the increase of the dimensionality of the data [4]. Many algorithms have been proposed to accelerate the k-means. In, the use of kd-trees [5] is suggested to accelerate the k-means. However, backtracking is required, a case in which the computation complexity is increased Clustering could also be defined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than individual data point. Basically, clustering is justify as a method used for grouping a large variety of knowledge into significant teams or clusters supported some similarity between data [6].

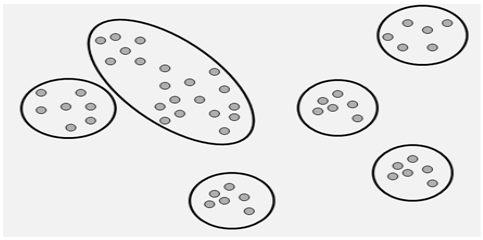

Figure1 Clustering

1. Partitioning Clustering

The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning and then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another [7]. In a good partitioning the objects in the same cluster are close or related to each other whereas objects in different clusters are far apart or different. Most applications adopt popular heuristic methods such as greedy approaches like the k-means and k-medoids algorithms which progressively improve the clustering quality and approach a local optimum

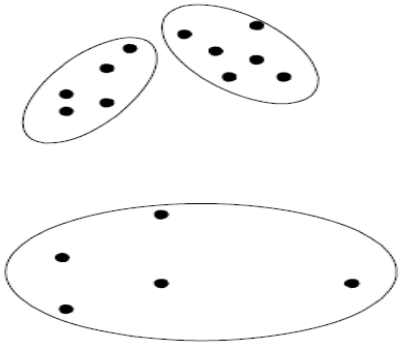

Figure2 Partitioning Clustering

2. Hierarchical Clustering:

A collection of nested clusters organized as a hierarchical tree hierarchical algorithms decompose a information D of n objects

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 3, Issue 10, October 2016)

into many levels of nested partitioning till every set consists of just one object. There are two varieties of hierarchical algorithms; an agglomerative that builds the tree from the leaf nodes up, whereas a divisive builds the tree from the highest down [3].

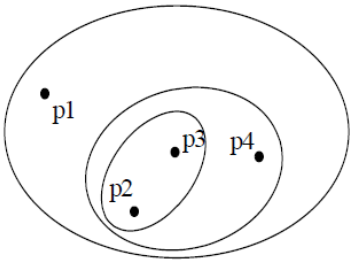

Figure3 Hierarchical Clustering

3. Applications of K-Mean Clustering

a. it is used for choosing color palettes on old fashioned graphical display devices and Image Quantization
b. It is relatively efficient and fast.
c. it is used on acoustic data in speech understanding to convert waveforms into one of k categories.
d. k-means clustering can be applied to machine learning or data mining.

4. Different Types of Clusters

a. Centre based clusters: A cluster is a group of objects so that an object in a cluster is more closer to the centrel 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 representativel point of a cluster [8].
b. Contiguous clusters: 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.
c. 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
d. 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 [8].

II. LITERATURE SURVEY

Amar Singh et al. [9]. They explained about Re-engineering software system is the recovery of software architecture and in software architecture recovery involves clustering. In this paper they guide us to introduce an approach that collectively clustering with matching technique to discover a decomposition which is well understood. Pattern matching is a technique under which architectural clues can be identified. All these clues are helpful to access an interclass similarity measure in clustering algorithm to produce the decomposition which is also known as final system decomposition. Adding a new updating in current existing software is always a challenging task but it also helpful to reduce the complexity in work. It is also necessary to keep update every error, patch or hack, for batter performance of any software system or a software architecture. Architectural clue collect the source model is designed with proper information.

Wang Shunye et al.[10] . A proposed a title "An Improved innovative Center Using K-means Clustering Algorithm and FCM"by the problem of random selection of initial centroid and similarity measures, the researcher presented a new K-means clustering algorithm based on dissimilarity. This improved k-means clustering algorithm basically consists of 3steps.The first step discussed is the construction of the dissimilarity matrix i.e dm.Secondly, Huffman tree based on the Huffman algorithm is created according to dissimilarity matrix. The output of Huffman tree gives the initial centroid. Lastly the k-means algorithm is applies to initial centroids to get k cluster as output. Iris, Wine and Balance Scale datasets are selected from UIC machine learning repository to test the proposed algorithm. Compared to traditional k-means the proposed algorithm gives better accuracy rates and results.

Navneet Kaur et al. [11.] 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..

Don Kulasiri et al. [12] .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.Nasim Akhtar et al [13] .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. sort is applied to sort the output that was previously generated. The data points are then divided into

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 3, Issue 10, October 2016)

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.

Juntao Wang et al. [14] .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. Rana et al. [15]. Proposed a 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.

Feng Xie et al. [16]. 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.

Jianchao Fan et al. [17]. 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.

Chetna Sethi et al. [18]. 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.

Fernandez et al. [19]. 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.

Ji Zhou et al. [20].They explained about the reverse engineering concept is quite famous these days and related to recovery of software architecture. There are number of technique which as used in this paper to recover software architecture, one of them is clustering technique, which source the same component from software. Generally the component feature is vague. A group of same data element is known as clustering. This technique is as older and it's used also in science and engineering. In simple words, identifying the number of data element, calculating similar coefficient and following the clustering method is called as clustering technique. The main function of the clustering technique is speedy and efficient recovery of software architecture by using fuzzy clustering technique. In this paper the major impact of this study shown that architecture recovery can be done batter by fuzzy clustering instead of ordinary clustering**.**

III. EXPECT OUTCOME

A study new research in the field of data mining and identify various challenges in the field of following objective to work in the field of center using clustering technique.

1. Minimum cluster and get useful information
2. Increase accuracy in clustering technique.
3. Find reliable data and best possible answer
4. Minimize error-values in clustering

IV. CONCLUSION

Clustering has a crucial role in different applications. The commonly used efficient clustering algorithm is k-means clustering. K-means clustering is an important topic of research now days in data mining. This paper has presented a survey of most recent research work done in this area. However k-means is still at the stage of exploration and development. The survey 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 we study the three clustering algorithms one is, simple K-Means partitioning

algorithm**.** and the GAKM an hybrid algorithm which is the combination of simple K-Means and Genetic Algorithm. K-means is combine with GA 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 genomic data set, numerical data set.

REFERENCES

[1]. Lv T., Huang S., Zhang X., and Wang Z, "Combining Multiple Clustering Methods Based on Core Group". Proceedings of the Second International Conference on Semantics, Knowledge and Grid (SKG'06), pp: 29-29, 2006.

[2]. Xu R., and Wunsch D., "Survey of clustering algorithms". IEEE Trans. Neural Networks, 16 (3): 645-678, 2005.

[3]. MacQueen J., "Some methods for classification and analysis of multivariate observations". Proc. 5th Berkeley Symp. Math. Stat. and Prob, pp: 281-97, 1967.

[4]. Kanungo T., Mount D.M., Netanyahu N., Piatko C., Silverman R., and Wu A.Y., "An efficient k-means clustering algorithm: Analysis and implementation". IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (7): 881-892, 2002.

[5]. Bentley J., "Multidimensional Binary Search Trees Used for Associative Searching. Commun". ACM,18 (9): 509-517, 1975

[6]. Sproull R., "Refinements to Nearest-Neighbor Searching in K-Dimensional Trees. Algorithmic", 6: 579-589, 1991.

[7]. Chih-Cheng Hung, Wenping Liu and Bor-Chen Kuo Marietta, "A new Adaptive fuzzy Clustering algorithm for remotely sensed images", GA 30060 USA ,2009.

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

[9]. Amar Singh and Navot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm", International journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2012.

[10]. Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity",2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22, Shenyang, China IEEE, 2013.

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

[12]. 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.

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

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

[15]. 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.

[16]. 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, pp. 1215-1218. 2002.

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

[18]. 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.

[19]. Fernández,Britos,Rossi,&GarcíaMartínezR.,"GeneticAlgorithmBasedImagecompression",SBAISimpósioBrasileiro de AutomaçãoInteligente, São Paulo, SP, 08-10 de Setembro de 1999.

[20]. Lingming Zhang, Ji Zhou, Dan Hao ,Lu Zhang, Hong Mei, "Prioritizing JUnit Test Cases in Absence of Coverage Information",IEEE,2009.