Current Trends in Technology and Science
ISSN: 2279-0535. Volume: 5, Issue: 6

Improved Innovative Center Using K-means Clustering Algorithm and EFCA

Arvind Dangi
M. Tech Scholar, Department of Computer Science & Engg Sam College of Engineering & Technology, Bhopal, M.P., India
Prof. Lokesh Malviya
Department of Computer Science & Engg, Sam College of Engineering & Technology, Bhopal, M.P., India.
lokesh.031986@gmail.com

Abstract — Data Mining is justify technique used to extract, which means full information from mountain information and clustering is a crucial task in data mininging process which can be used for the aim to make groups or clusters of the particular given information set that's predicated on the similarity between them. K-Means cluster may well be a cluster procedure throughout that the given info set is split into K i.e. type of clusters. The impact issue of k-means is its simplicity, high efficiency and quality. However, is additionally contains of type of limitations: random selection of initial centroids, type of cluster K got to be initialized and influence by outliers. visible of these deficiencies, our planned approach of an Improved innovative Center using K-means cluster rule and proposed algorithm enhancements to traditional k-means to handle such limitations which we are able to compare K-means clump rule with varied clump rule. Increase accuracy of the perform cluster new technique are going to be planned to with efficiency cluster the functions consistent with their importance.

Keyword — Data mining, clustering algorithm, EFCA

1. INTRODUCTION

The field of data mining and information discovery is up-and-coming as a new, basic study area with important applications to Science, engineering, medicine, business, and education. Data mining attempts to formulate analysis and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data [1]. Size of databases in scientific and commercial application is huge where the number of records in a dataset can vary from some thousand to thousand of millions .Clustering may be defined as a data reduction tool i.e. used to create subgroups that are more and more manageable than individual datum. Basically, clustering is justify as a process used for grouping a large number of information into groups or clusters in some objects similarity between data. Cluster is the groups that have data similar on basis of common features and dissimilar to data in other clusters. Data Mining is the process of extracting hidden, previously unknown and useful information from large databases and data warehouses. Data mining process involve steps like data cleaning, addition, selection, change in data mining technique, pattern matching evaluation. Various data mining techniques are used like classification, clustering, relationship rules, in order patterns, Prediction, Decision trees, etc. different applications used. Here discuss about clustering algorithms like fuzzy c-means, k-means [2].

1.1TYPES OF CLUSTERS

1. 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. Centre based clusters: A cluster is a group of objects so that an object in a cluster is more 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. 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.

4. 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

1.2 APPLICATIONS OF K-MEAN CLUSTERING

1. It is relatively efficient and fast. It computes result at O (tkn), where n is number of objects or points, k is number of clusters and t is number of iterations.

2. k-means clustering can be applied to machine learning or data mining

3. Used on audio information in speech understanding to exchange waveforms into one of n categories

4. Also used for choosing color palettes on old fashioned graphical display devices and Image Quantization.

2. LITERATURE SURVEY

Wang Shunye et al. [3]. Has 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.

Navjot Kaur et al.[4]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-hierarchical 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..

Yang et al. [5]. 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. S. Mahmud et al. [6]. 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. sort 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.

Juntao Wang et al. [7] .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 Likes 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. [8] 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.

Feng Xie et al. [9] 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. [10] 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.

K. Premalatha et al. [11] presented the hybrid approach of PSO with Genetic Algorithm (GA). The proposed hybrid PSO systems find a better solution without trapping in local maximum, and to achieve faster convergence rate. This is because when the stagnation of PSO occurs, GA diversifies the particle position even though the solution is worse. This makes PSO-GA more flexible and robust. Unlike standard PSO, PSO-GA is more reliable in giving better quality solutions with reasonable computational time. Experiment results are examined with benchmark functions and results show that the proposed hybrid models outperform the standard PSO.

Chetna Sethi et al. [12] 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.

Ahmed A. A. Esmin et al. [13] presented a literature survey on the PSO algorithm and its variants to clustering high dimensional data. An attempt is made to provide a guide for the researchers who are working in the area of PSO and high dimensional data clustering.

V. Katari et al. [14]. An improved genetic algorithm (IGA) was proposed in which an efficient method of crossover and mutation were implemented. The proposed algorithm was a combination of GA, the popular Nelder-Mead (NM) Simplex search and K-means to find optimal solution.

Han-Saem Park et al. [15].GA based clustering techniques have a large area of application. A few of these applications are discussed in this paper. An evolutionary fuzzy clustering method with knowledge-based evaluation was proposed in to identify unknown functions of genes.

Merlo et al. [16]. The image compression problem using genetic clustering algorithms based on the pixels of the image was proposed. GA was used to obtain an ordered representation of the image and then the clustering was performed to obtain the compression.

A.M. Natarajan et al. [17]. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. The clustering algorithm attempts to find natural groups of components, based on some similarity. Traditional clustering algorithms will search only a small sub-set of all possible clustering and consequently, there is no guarantee that the solution found will be optimal. This paper presents the document clustering based on Genetic algorithm with Simultaneous mutation operator and Ranked mutation rate. The mutation operation is significant to the success of genetic algorithms since it expands the search directions and avoids convergence to local optima. In each stage of the genetic process in a problem, may involve aptly different mutation operators for best results. In simultaneous mutation the genetic algorithm concurrently uses several mutation operators in producing the next generation. The mutation ratio of each operator changes according to assessment from the respective offspring it produces. In ranked scheme, it adapts the mutation rate on the chromosome based on the fitness rank of the earlier population. Experiments results are examined with document corpus. It demonstrates that the proposed algorithm statistically outperforms the Simple GA and K-Means.

3. Simulation and Result Analysis

Environment: 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.

Performance comparison between K-mean & proposed algorithm

Clustering Technique	Dataset	Time (in Sec)	Error Rate
K-means Algorithm	Lang cancer dataset	3.05762	4.69069
Proposed Algorithm		5.74084	1.93009
K-means Algorithm	e_coili dataset	1.98121	3.09229
Proposed Algorithm		1.90321	0.445803

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 they use Intel 1.4 GHz Machine and operating system window7, window-xp etc. MATLAB version 14 (R2008a) is a high-level technical compute language**.**

1. Lung Cancer Dataset

Lung cancer dataset used and set threshold value .777 and apply k-mean algorithm and EFC algorithm .k-mean algorithm is more error rate as compare to proposed algorithm. But k-mean algorithm time takes minim as compare to proposed algorithm**.** Proposed algorithm is better as compare to k-means because k-mean data redundancy is more but P proposed algorithm is minim redundancy .it is show figure1, below graph show time is more but cluster error rate minimum also called minimize redundancy in dataset. Proposed algorithm is get fine data in lung cancer dataset. Proposed algorithm.

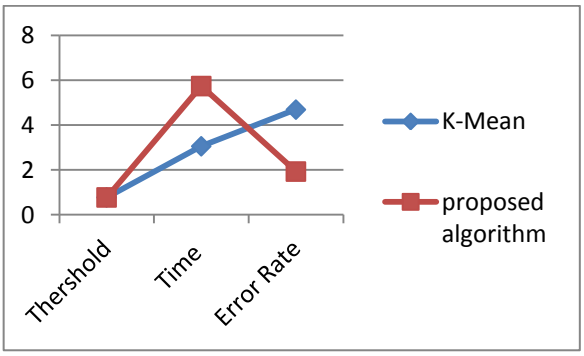


Figure 1 Performance analysis between K-mean & proposed algorithm

2. E. coli Dataset:

Name is also called Escherichia coli. E. coli Dataset used and set threshold value .011 and apply k-mean algorithm and proposed algorithm. K-mean algorithm is more error rate as compare to proposed algorithm .but k-mean algorithm time take minim as compare to proposed algorithm. Proposed algorithm is better as compare to k-means because k-mean data redundancy is more but proposed algorithm is minim redundancy .it is show figure2, below graph show time is more but cluster error rate minimum also called minimize redundancy in dataset. Proposed algorithm is get fine data in E. coli Dataset

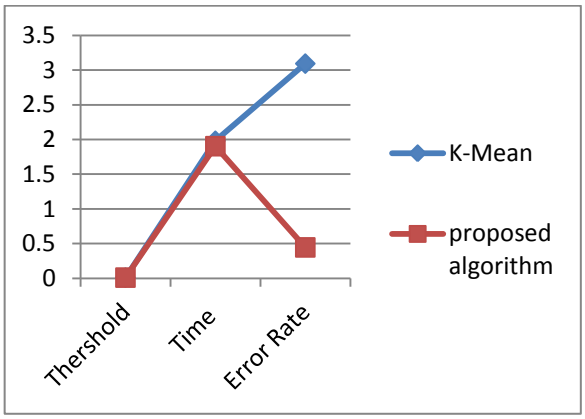
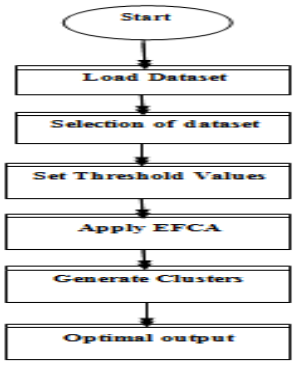


Figure 2 Performance analysis between K-mean & proposed algorithm

Working of proposed algorithm:



4. CONCLUSION

To simulate an improved innovative Center using proposed algorithm and k-mean algorithm used totally different datasets, in experimentation each the algorithms for used totally different datasets with famous clustering algorithms, lung cancer dataset and E_coli dataset. In order that clustering algorithm with higher accuracy is analysis with totally different cluster groups. It might speed up clustering method. Choose randomly initial center for the k-means algorithm, Experimental results show that improved initialization center algorithm for Enhance fuzzy clustering algorithm greatly will increase the quality and stability of the algorithm. Enhance fuzzy clustering algorithm is better as compare to k-mean algorithm because more accuracy supported minim error rate. In future will be improving K-means algorithm. In future analysis are minimizing the execution time of proposed algorithm. Within the method of minimizing of iteration of proposed algorithm. And also improving accuracy of k-mean algorithm.

References

[1]. Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 121-125, 2010.

[2] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1, London, UK, July 2009.

[3] 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.

[4] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur "Efficient K-means 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.

[5] 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.

[6] 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 December, 2012, Dhaka, Bangladesh, IEEE , 2012.

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

[8] 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.

[9] 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.

[10] 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 , pp. 1-12, 2013.

[11] K. Premalatha and A.M. Natarajan, "Hybrid PSO and GA for Global Maximization," ICSRS, Int. J. Open Problems Compt. Math., Vol. 2, No. 4, , pp. 597-608, December 2009.

[12] 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

[13] Ahmed A. A. Esmin, Rodrigo A. Coelho and Stan Matwin, "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data," Springer, pp.-1-23, Feb 2013.

[14] V. Katari, S. C. Satapathy, JVR Murthy,P. Reddy ,"A Hybridized Improved Genetic Algorithm with Variable LengthChromosome for Image Clustering", International Journal ofComputer Science and Network Security, VOL.7 No.11,November 2007.

[15]Han-Saem Park, Si-Ho Yoo, and Sung-Bae Cho "Evolutionary Fuzzy Clustering Algorithm with Knowledge-Based Evaluation and Applications for Gene Expression",Journal of Computational and Theoretical NanoscienceVol.2,1–10, 2005

[16] Merlo, Caram, Fernández, Britos, Rossi, &GarcíaMartínezR.,"Genetic-Algorithm Based ImageCompression",SBAI–SimpósioBrasileiro de AutomaçãoInteligente, São Paulo, SP, 08-10 de Setembro de 1999.

[17] K. Premalatha, A.M. Natarajan, "Genetic Algorithm for Document Clustering with Simultaneous and Ranked Mutation", Modern Applied Science, Vol. 3, No. 2, February, 2009.