International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 02, Issue: 03 (MAY-JUNE, 2016)

# Improved Innovative Center of K-means Clustering Algorithm via FCM

**Arvind Dangi**
M. Tech., Scholar Department of CSE
Sam College of Engineering & Technology, India
arvinddangi3@gmail.com

**Prof. Pankaj Singh**
Department of CSE
Sam College of Engineering & Technology, India
pankajonline2u@gmail.com

***Abstract--*** Data Mining is justify technique used to extract, which means full information from mountain information (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 victimization K-means cluster rule and FCM 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.

***Keywords:*** **-** K-Mean, FCM, PSO, Genetic Algorithm, medoid algorithm.

## I. INTRODUCTION

The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to Science, engineering, medicine, business, and education. Data mining attempts to formulate analyse and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data [8]. 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 [7]. 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 data into meaningful groups or clusters based on some similarity between data. Clusters are 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, integration, selection, transformation, data mining technique, pattern evaluation and knowledge representation. Various data mining techniques are used like classification, clustering, association rules, Sequential patterns, Prediction, Decision trees, etc. are used in various applications. Here we discuss about clustering algorithms like fuzzy c-means, k-means.

### 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 center 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 acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization or Image Segmentation).

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

## II. LITERATURE SURVEY

Wang Shunye [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, Navneet Kaur [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 analyzed 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 [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. Sohrab Mahmud [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 [9] 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 centers. In the next step fast global k-means algorithm proposed by Aristides Likes is applied to the output generated previously. The results between k-means and improved k-means are compared using Iris, Wine, and Abalone 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. [10] 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. [11] 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. [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.

K. Premalatha et al. [13] 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. [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.

Ahmed A. A. Esmin et al. [15] 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.

An improved genetic algorithm (IGA) was proposed in [16] 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 Kmeans to find optimal solution.

A genetic based clustering algorithm called GASDCA (GA with point Symmetry Distance based Clustering Algorithm) was proposed in [17] which was able to detect both convex and non convex clusters. The proposed GASDCA was compared with existing symmetry based clustering technique, SBKM, its modified version, Mod-SBKM and the well known Kmeans algorithm [23]

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 [18] to identify unknown functions of genes. The image compression problem using genetic clustering algorithms based on the pixels of the image was proposed in [19]. GA was used to obtain an ordered representation of the image and then the clustering was performed to obtain the compression. A GA was proposed in [20] to deal with document clustering. The GA algorithm calculated the optimum value of k and solved the best grouping of the documents into these k clusters. The performance of this algorithm was evaluated on datasets of documents that were the output of a query in a search engine.

## III. EXPECT OUTCOME

We have study newly research paper in the field of data mining and identify various challenge so in this synopsis our objective to want the challenges in the field of following objective to work in the field of an improved innovative center using K-means clustering algorithm and FCM technique.

1. An increase accuracy of the function clustering new technique.
2. A efficiently cluster the functions according to their importance.
3. A need to be careful as increasing k results in smaller error-function values by definition.
4. A come near is to compare the results of multiple runs with different k clusters and choose the best one according to a given criterion

## IV. CONCLUSION

We want simulate an improved innovative Center Using K-means Clustering Algorithm and for different datasets, so that clustering with better accuracy is analysed with different cluster groups. When considering datasets of humans and different species, this improvement becomes very useful, because analysis of huge amount of gene expression has been made easy, as the datasets of all organisms have large amounts of data or sample. When examining traditional K-means algorithm, traditional K-means algorithm lacks in performance or in choice of initial clustering centre. The algorithm has been shown to be very effective in clustering multidimensional data sets. The algorithm has been tested on synthetic data and iris dataset, with different datasets.

## 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 , PP. , 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, Proceedings of the World Congress on Engineering, Volume 1, July 2009.

[3]. Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity" ,International Conference on Mechatronic Sciences, Electric Engineering and Computer, China IEEE , Dec. 2013.

[4]. Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur "Efficient Kmeans clustering Algorithm Using Ranking Method In Data Mining" ,International Journal of Advanced Research in Computer Engineering & Technology, ISSN: 2278–1323, Vol. 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,Issue 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 Dhaka, Bangladesh, IEEE, December 2012.

[7]. Birdsall, C.K., Landon, A.B Plasma Physics via Computer Simulation, Adam Hilger, Bristol, 1991.

[8]. M H Dunham, "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2002.

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

[10]. 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.

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

[12]. 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-2013-0157.

[13]. K. Premalatha and A.M. Natarajan, "Hybrid PSO and GA for Global Maximization," ICSRS, International Journal Open Problems Computer Mathematics, Vol. 2, No. 4, PP. 597-608, December 2009.

[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]. 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.

[16]. V. Katari, S. C. Satapathy, JVR Murthy,P. Reddy ,"A Hybridized Improved Genetic Algorithm with Variable Length Chromosome for Image Clustering", International Journal of Computer Science and Network Security, Vol. 7 No.11,November 2007.

[17]. S. Saha, S. Bandyopadhyay, U. Maulik," A New Symmetry-Based Genetic Clustering Algorithm", Machine Intelligence Unit, Indian Statistical Institute, India.

[18]. 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

[19]. Merlo, Caram, Fernández, Britos, Rossi, &Garcia Martinez R., "Genetic Algorithm Based Image Compression", SBAI–Symposia Brasileiro de Auto Macao Intelligent, São Paulo, PP, 08-10, September 1999.

[20]. A. Casillas, M. T. Gonzalez de Lena, and R. Martınez, "Document Clustering into an unknown number of clusters using a Genetic Algorithm".