International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 04, Issue: 03 (May- June, 2018)

An Efficient Approach to Enhance Performance of FCM Using PFA

Ashwani Kumar [1], Prof. B.P.S. Senger[2]
M. Tech. Scholar Department of CSE
ASCE, RGPV, Bhopal, M.P., India
[1]ashwani.mskjuat@gmail.com, [2]bpssenger@gmail.com

ABSTRACT- *Data mining organizes the wide dataset and variety of data sets automatically and acquires accurate classification Method. The fuzzy c-means is a frequently utilized algorithm at present. Sensitive to the initial number and centers of clusters is one shortcoming of fuzzy c-means clustering method. It is very popular clustering algorithm also called fuzzy c-means algorithm (FCM) and the problem of usual dataset duplicate sets to perform feature fuzzy clustering. Clusters that have been wrongly separated resulting in some clusters close to each other. The close clusters can be found by investigating the partition matrix. Those close clusters should be divided or merged. Proposed new method to update the appropriate clusters and cluster centers. The unsupervised techniques like clustering techniques are very much suitable for handling UCI dataset in this case the learning parameters are computed from learning data. Using clustering algorithm they can assign labels to unlabeled dataset and reduce similarity between different clusters. Proposed methods based on particle swarm optimization. Improvement centers in the cluster Initial cluster centers and the local optimum every cluster. New method particle swarm optimization with fuzzy c-means algorithm (PFA) uses as cluster validity and finds the optimal cluster centers. PFA is base on particle values minimize in cluster element number within a specific range with cluster partitions that provide compact and well-separated clusters. Experiments show that the proposed approach significantly improves the clustering effect. Experiments on synthetic datasets and a real dataset show that the proposed clustering method has good performance by comparing to the standard fuzzy c-means clustering method.*

Keywords:- *Data Mining, Partition Clustering, Clustering algorithms, Fuzzy Clustering, FCM Algorithm.*

I. INTRODUCTION

Data mining has been verified mutually of the favorite areas of researchers with its completely outstanding applications in type of fields, exploring new approaches to see, examine, and uncover any hidden patterns in information automatically. Data processing system may be outlined as an integration of techniques from multiple disciplines specifically statistics, info warehouse, info retrieval, machine learning, pattern recognition, neural networks, image & signal process, and computing etc. In information technology driven society, wherever data is a useful plus to anyone, organization or government Firms are provided with large quantity of information in day after day, and there's the necessity for them to concentrate on processing these knowledge thus on get the foremost vital and helpful info in their information warehouses. Data processing may be a new technology that may be employed in extracting valuable info from information warehouses and databases of firms and governments. It involves the extraction of hidden info from some immense dataset. It helps in detection anomalies in information and predicting future patterns and perspective in a very extremely economical manner. Applying data processing makes it easier for firms and government, throughout quality selections from on the market information, which might have taken longer time, supported human experience. Data {processing} is that the process of extracting helpful and hidden info or data from information sets. The data thus extracted may be accustomed improve the choice creating capabilities of an organization or a company [1].Data mining consists of six basic varieties of tasks that are Anomaly detection, Association rule learning, Clustering, Classification, Regression and report. Cluster is one among the vital tasks of information mining. Cluster is that the unsupervised classification of information objects into teams or clusters. Cluster is outlined because the task of grouping objects in such how that the objects within the same group/cluster share some similar properties/traits. Several non-GA-based algorithms like K-means and Fuzzy-c-means have been used for the clump tasks. One among the most goals of cluster algorithms is to search out „natural" teams within the dataset together with partitioning the info into those natural teams. However none of those non-GA-cluster algorithms are efficient enough to get „natural" teams from all the input patterns, especially when the number of clusters included in the data set tends to be large. These algorithms also suffer from the problem of local convergence due to large clustering search space [2]

Data mining consists of major elements Classification

1. Clustering: Clustering is that the illustration of information in categories. However, not like classification, in clustering, category labels are unknown and it's up to the clustering rule to get acceptable categories. This known as unsupervised classification clustering could be a collection of information objects, similar information are taking within the same cluster, dissimilar information are taking in numerous clusters. Cluster is that the method of assignment a homogeneous cluster of objects into subsets known as clusters, in order that objects in every cluster are additional like one another than objects from totally different clusters supported the values of their attributes. Cluster techniques are studied extensively in data processing, pattern recognition and machine learning. Cluster algorithms are often usually sorted into 2 main categories, namely, supervised cluster and unsupervised cluster wherever the parameters of classifier are

optimized. Several unsupervised agglomeration algorithms are developed. One such algorithms *k*-means, that assigns *n* objects to *k* clusters by minimizing the ad of square Euclidian distance between the objects in every cluster to the cluster center. The most downside of the *k*-means rule is that the result's sensitive to the choice of initial cluster centroids and should converge to local optima [3].

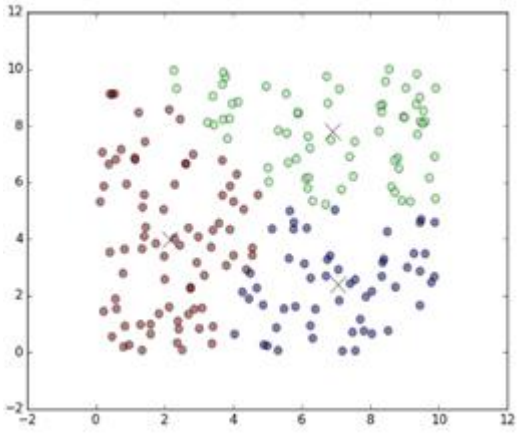

Figure1 Different Clusters

2. Association: Association analysis is that the discovery of association rules. It depends on the frequency of transactional information occur along in info, additionally depends on a threshold referred to as support, and identifies the frequent item sets. Association data processing aimed to search out association between attributes, generate rules from information sets. The association rule mining role is to succeed in all rules having support minus (minimum support) threshold and confidence minconf (minimum confidence) threshold [4]. Several papers are dedicated to develop algorithms to mine standard association rules. The first efficient algorithms like Apriori and Apriority, SETM, OCD, and DHP were continuing with newer developments like DIC, CARMA, TBAR and FP-Growth.

(i) Fuzzy association rules: Association rules are often integrated with several techniques i.e. fuzzy rules

(a) Crisp rule: Crisp pure mathematics uses one among only 2 values: true or false. Crisp set cannot represent obscure ideas. Components are assigned to the sets by giving them the values zero or one. Each component with one worth may be a member of the set, components with zero worth is non-member of the set. The quantity of components that belong to a group is named its cardinality.

(b) Fuzzy rule: Fuzzy sets described as an extension of the classical crisp sets. Fuzzy pure mathematics is that part belongs to a fuzzy set with an exact degree of membership. Thus, a proposition isn't either true or false, however could also be part true or part false to any degree. This degree is typically taken as a true variety within the interval [0, 1]. Fuzzy rules are often combined with Association rules to get fuzzy association rules. There exists a replacement approach that use classical association rule mining by victimization fuzzy sets. Fuzzy association mining solves several issues notice in large quantities of information that exist usually in info with efficiency. Once dividing an attribute within the information into sets covering sure ranges of values, we tend to are confronted with the sharp boundary drawback. Components close to the boundaries of a crisp set can either be unheeded [5].

3. Neural Networks in Data Mining: Neural networks are with success applied in supervised and unsupervised learning applications .There are 2 categories of approaches for data processing with neural networks .The first approach referred to as rule extraction involves model extraction from trained neural networks, The second approach is to directly learn easy easy-to-understand networks .Neural Network Applications will be classified in following categories [6].

(i) Clustering: A cluster algorithmic rule explores the similarity between patterns and places similar patterns in a very cluster. Best best-known applications embrace information compression and data processing.

(ii)Classification/Pattern recognition: The task of pattern recognition is to assign an input pattern (like written symbol) to 1 of the many categories. This class includes recursive implementations like associative memory.

(iii) Perform approximation: The tasks of perform approximation is to search out an estimate of the unknown perform f () subject to noise. varied engineering and scientific disciplines need perform approximation.

(iv) Prediction/Dynamical Systems: The task is to forecast some future values of a time sequenced information. Prediction includes a vital impact on call support systems. Prediction differs from perform approximation by considering time issue. Here the system is dynamic and should manufacture totally different results for identical computer file supported system state (time).

Data processing Applications

There are several applications in data processing [7]

(i)Medical information mining: Over the past decade, nudged by new federal laws, hospitals and medical offices round the country are changing scribbled doctors' notes to electronic records. Though the chief goal has been to enhance potency and cut prices.

(ii)Spatial data processing: spatial information mining is that the application of knowledge mining strategies to spatial data, the end objective of spatial data processing is

to search out patterns in knowledge with reference to earth science. So far, data processing and Geographic data Systems (GIS) have existed as 2 separate technologies, every with its own strategies, traditions, and approaches to image and knowledge analysis. Significantly, most modern GIS have only terribly basic spatial analysis practicality. The huge explosion in geographically documented information occasioned by developments in IT, digital mapping, remote sensing, and therefore the world diffusion of GIS emphasizes the importance of developing information driven inductive approaches to geographical analysis and modeling .

II. RELATED WORK

Marty et.al [8]. In the paper, however cluster technique is helpful to identify totally different data by considering varied examples and one will see wherever the similarities and ranges agree. By examining one or a lot of attributes or categories, you'll cluster individual items of information along to make a structure opinion. At an easy level, cluster is victimization one or a lot of attributes as your basis for identifying a cluster of correlating results. cluster will work each ways that. you'll assume that there's a cluster at sure purpose and so use our identification criteria to see if you're correct.

K. A. Abdul Nazeer et al [9].the major disadvantage of the k-means algorithmic rule is regarding choosing of initial centroids that produces totally different clusters. However final cluster quality in algorithmic rule depends on the choice of initial centroids. 2 phases includes in original k means that algorithm: 1st for determinative initial centroids and second for distribution information points to the closest clusters then recalculating the cluster mean. However this increased cluster technique uses each the phases of the first k-means algorithmic rule. This algorithmic rule combines a scientific technique for locating initial centroids and an efficient means for distribution information points to clusters. However still there's a limitation during this increased algorithmic rule that's the worth of k, the amount of desired clusters, continues to be needed to be given as an input, despite the distribution of the information points.

Hart P et al. [10]. Describes the varied uses of ordered patterns for identifying trends, or regular occurrences of comparable events. as an example, with client knowledge you'll determine that customers get a specific collection of product along at completely different times of the year. During a shopping basket application, you'll use this data to automatically counsel that bound things be additional to a basket supported their frequency and past buying history.

Soumi Ghosh et al. [11] proposed a comparative discussion of 2 cluster algorithms particularly centroid based mostly} K-Means and representative object based Fuzzy C-Means cluster algorithms. This discussion is on the premise of performance analysis of the potency of cluster output by applying these algorithms. The results of this comparative study is that FCM produces nearer result to the K-means however still computation time is over k-means because of involvement of the fuzzy measure calculations.

Wang Shunye et al. [12]. A planned a title "An Improved innovative Center victimization K-means cluster rule and FCM" by the matter of random choice of initial centroid and similarity measures, the research worker conferred a replacement K-means cluster rule supported dissimilarity. This improved k-means cluster rule primarily consists of 3steps.The first step mentioned is that the construction of the dissimilarity matrix i.e. dm.Secondly, Huffman tree supported the Huffman rule is made per dissimilarity matrix. The output of Huffman tree provides the initial centroid. Last the k-means rule is applies to initial centroids to induce k cluster as output. Iris, Wine and Balance Scale datasets are chosen from UIC machine learning repository to check the planned rule. Compared to traditional k-means the planned rule provides higher accuracy rates and results.

Th. Shanta Kumar et al. [13]. Gave a comparative analysis between k-means cluster algorithmic rule and fuzzy cluster algorithmic rule. In this paper the scientist additionally discuss the benefits and limitations of fuzzy c-means algorithmic rule's-means may be a partional primarily {based} cluster algorithmic rule whereas Fuzzy cmeans is non partional based cluster algorithmic rule. Fuzzy cmeans primarily works in two methods. In the 1st method cluster centers area unit calculated and in second the information points are assigned to calculated cluster center with the assistance of geometer distance. This method is sort of the same as standard k-means with a bit distinction. In fuzzy cmeans algorithmic rule membership price starting from zero to one is assigned to information item in cluster.0 membership indicates that the information purpose isn't a member of cluster whereas one indicates the degree to that data point represents a cluster. The problem faced by fuzzy c-means algorithmic rule is that add of membership value of information points in every cluster is restricted to one. Algorithm additionally faces downside in handling outliers. On the opposite hand comparison with k-means shows that the fuzzy algorithmic rule is efficient in getting hidden patterns and information and knowledge and information from natural data with outlier points.

Sijia Liu et al. [14] described a helpful survey of fuzzy clustering in main 3 classes. The primary class is essentially the fuzzy cluster depends on exact fuzzy relation. The second is that the fuzzy cluster supported single objective performs. Finally, it's given an outline of a statistic classifier. That's the fuzzy generalized k nearest neighbor rule. The fuzzy cluster algorithms have obtained nice success in an exceedingly kind of substantive areas.

Hui-ping et.al. [15]. Health care has been the foremost apace growing phase. Thyroid is an ductless gland that is one in every of the foremost vital organs that sounds like a butterfly and it's set at the lower a part of the neck within the body. Thyroid is chargeable for producing the thyroid hormones that is responsible for dominant the metabolism. These hormones are unleash into the blood and carried to the tissue via blood. It regulates the body functions like energy usage, dominant the temperature and keepings organs operating as desired. two main thyroid hormones secreted by the ductless gland are T (T3) and levothyroxine (T4). These hormones have an effect on nearly all tissues of the body and will increase cellular activity. The thyroid lobe is prostrate to many terribly distinct troubles, as an example, goitres, thyroid cancer, solitary thyroid nodules, thyrotoxicosis, glandular disease, and thyroiditis.

III SIMULATION ENVIRONMENT AND RESULT ANALYSIS

(a) Proposed new technique particle swarm optimization with fuzzy c-means algorithm (PFA) uses as cluster validity and finds the optimal cluster centers. PFA is base on particle values minimize in cluster element number within a specific range with cluster partitions that provide compact and well-separated clusters. Data mining clustering with proposed algorithms PFA have recently been shown to find good results in a wide variety of real-world dataset. The variety of variations in proposed algorithms PFA based clustering techniques are proposed algorithms in achieving better in particular multidimensional dataset. PFA technique using dataset and generate different clustering predicting the primary cluster center scene and number of clusters and optimizing it with one of efficient cluster. The proposed approach is intended to design a system that will overcome the limitations associated with clustering where multidimensional data is of main concern. Hence, the proposed algorithm will resolve dead unit problem, stagnation, multiple cluster membership as well as premature convergence to local optima.

(b) A MATLAB (2013a) is a software language initially developed by Math Works for numerical and mathematical computations also for symbolical manipulations. To opening it just double clicks the MATLAB (2013a) browser. An MATLAB (2013a) environment will open as shown in the Figure A.1. This contains three basic windows, containing a large Command Window at the right, Working memory or Workspace and Command history windows on the left hand side. At the Command Window all the calculations are carried out in MATLAB (2013a). Another small window display information about the current MATLAB (2013a) directory or session, and our computer account. These small windows are named as Command History. This basically displays the commands typed in both current and previous sessions. Current Directory, The Workspace, which displays information about all output variables defined in current session. The Figure explains the basic MATLAB working environment. It is the high level language and interactive background used by millions of engineers and scientists universal. It lets explore and visualize ideas and work together across different disciplines with signal and image processing, message and calculation of results. MATLAB (2013a) provides implements to obtain, analyze, and picture data, allow 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 (2013a) code. Matrix laboratory is a multi paradigm numerical compute condition and 4th invention programming language. It is developed by math work; MATLAB (2013a) allows matrix strategy, plotting of function and data, implementation of algorithm, construction of user interfaces with programs.

(c) UCI Dataset Analysis: experiment perform four medical data like Thyroid_Dataset, IRIS_Dataset, BCWO_Dataset and ETEC_Dataset .two clustering algorithm FCM and PFA are compare result in table 6.1.this experiment time using set threshold point values like (p1,p2,p3….) and so on. here set threshold value p1=.485.

Table 1 Comparing the Performance of FCM and PFA Algorithms

Dataset	Algorithm	Execution Time (in sec)	Error Rate (in db)
Thyroid_Dataset	FCM	2.88602	3.99423
	PFA	1.93441	1.30723
IRIS_Dataset	FCM	2.33922	4.57657
	PFA	2.35562	1.76781
BCWO_Dataset	FCM	2.73002	4.47557
	PFA	2.38682	1.67781
ETEC_Dataset	FCM	3.38522	5.1847
	PFA	2.91722	2.20803

UCI Dataset Graph Result Analysis:

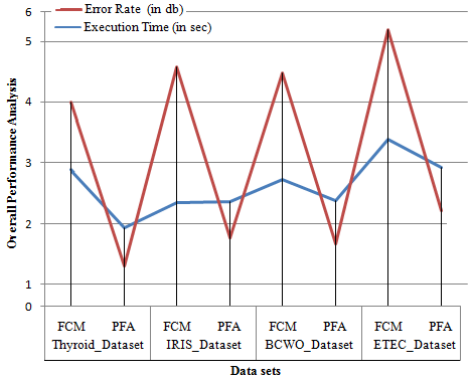

Figure3 Performance Analysis between FCM and propose algorithm

IV.CONCLUSION

Proposed technique particle swarm optimization with fuzzy c-means algorithm (PFA), it improved clustering uses as cluster validity and finds the optimal cluster centers an algorithm to avoid giving cluster number in advance and error reduce in cluster the sensitivity value. Close-cluster groups are found and classified to the types

of inappropriate clusters. Thus the adjustment is given as each type has its dividing or merging solution. Also, the approach to update cluster center is given to each type. Thus all the close-cluster groups can find the corresponding solution to update both cluster number and cluster centers with the updated cluster centers as labeled patterns. The proposed algorithm is projected to exercise on multidimensional medical data sets to achieve improved clustering results showing expected performance by decreasing the error rate and average the time complexity but increase accuracy of dataset. The experiments carried on both synthetic data and real data prove that the algorithm can get the appropriate clustering result with initial cluster number changing in a certain range. Proposed so far have been applied to different types of datasets, small data sets as well as large data sets, simple datasets as well as multivariate datasets. The FCM and PFA have same runtime complexity but slow when compared with other clustering algorithms Based on the primary factors like execution time and cluster quality. Find and concludes that many improvements are basically required on particle swarm optimization with fuzzy c-means algorithm to improve problem of cluster initialization, cluster quality and error less data of our propose algorithm**.** Enhance Clustering Algorithm based on PFA Clustering get the optimize number of clusters. Both algorithm are simple to understand and can be applicable for various type of dataset. Minimum fault values in dataset and find useful data values and increase accuracy in clustering get reliable data.

REFERENCES

[1]. A. A. Freitas, "Data Mining and Knowledge Discovery with Evolutionary Algorithms", Springer, 2002.

[2]. K. Krishna and M. N. Murty,"Genetic K-Means Algorithm", IEEE Transaction On Systems, Man, And Cybernetics Part B: Cybernetics, Vol. 29, No. 3, June 1999.

[3]. H.-P. Kriegel, P. Kroger, and A. Zimek, "Clustering high dimensional data: a survey on subspace clustering, pattern based clustering, and correlation clustering," ACM Transactions on Knowledge Discovery from Data, vol. 3, no. 1, article 1, 2009.

[4]. Jiawei Han, "Data Mining - Concepts and Techniques", 2nd Edition –Impression, 2006.

[5]. Bakk. Lukas Helm, "Fuzzy Association Rules", an Implementation in R, 2007.

[6]. P. Dostál, P. Pokorný, "Cluster Analysis and Neural Network ", 2009.

[7]. Mrs. Bharati M. Ramager , "Data Mining Techniques And Applications", Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305,2010

[8]. BABU, G.P. and MARTY, M.N., "Clustering with evolution strategies Pattern Recognition", 27, 2, 321-329, 1994.

[9]. K. A. Abdul Nazeer, M. P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 VOL.1 2009, July 1 - 3, 2009, London, U.K.

[10]. DUDA, R. and HART, P., "Pattern Classification and Scene Analysis". John Wiley & Sons, New York, NY, 1973.

[11]. Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[12]. 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.

[13]. Raju G, Binu Thomas, Sonam Tobgay and Th. Shanta Kumar "Fuzzy Clustering Methods in Data Mining: A comparative Case Analysis" 2008 International Conference on advanced computer theory and engineering, 2008 IEEE.

[14]. Sijia Liu, Don Kulasiri, Philip K. Maini and Radek Erban, " 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.

[15]. Elpiniki, I., papageorgiou, Nikolaos I. papandrianos, Dimibios J. Apostolopoulos and Pavlos J. Vassilakos Fuzzy cognitive map based decision support system for the thyroid diagnosis management in IEEE conference on Fuzz, 2008.

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

[17]. Paolo Tonella, Paolo Avesani, Angelo Susi 22nd IEEE International Conference on Software Maintenance (ICSM'06), 2009.Using the Case-Based Ranking Methodology for Test Case Prioritization.

[18]. Min Wei, Tommy W. S. Chow and Rosa H. M. Chan, "Clustering Heterogeneous Data with K-means by Mutual Information-Based Unsupervised Feature Transformation". Entropy 2015, 17, 1535-1548.

[19]. Bhagyashree Pathak, Niranjan Lal, "A Survey on Clustering Methods in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 159 – No 2, February 2011.

[20]. Nidhi Singh, Divakar Singh, Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time, International Journal of Computer Science and Information Technologies, Vol. 3(3),4119-4121, 2012.