International Journal of Innovative Research in Technology & Science

Received 3 October 2018, Revised Received, 9 November 2018, Accepted 19 November, 2018, Available online 25 November , 2018

ISSN: 2321-1156 Volume VI Issue VI, October 2018

Improved Dataset Analysis Performance Based on FCM Clustering Algorithm and PVCA

Kanchan Pandey, Vindhya Institute of Technology and Science, ,Satna, M.P., India; kanchanp1150@gmail.com
Nilesh Shrivastava, Department of CSE, Vindhya Institute of Technology and Science, Satna, M.P.,

ABSTRACT

Data mining is that the method of collection and analyzing helpful patterns from large quantity of information, its major functions, and cluster is one among them. In cluster, they create clusters of same information and duplicate information in FCM. The things in one cluster of cluster are alike whereas completely different from items that square measure in another cluster of cluster. Cluster could be a task of assignment a group of objects into teams referred to as clusters. Generally the cluster algorithms are often classified into different classes. One is difficult cluster; another one is soft (fuzzy) clustering. Onerous clusters, the data's are divided into distinct clusters, wherever every information component belongs to precisely one cluster. In soft cluster, information parts belong to over one cluster, and related to every component could be a set of membership levels but, is to boot contains of variety of limitations, random choice of initial centroids. During this technique, the data is initial clustered with normal fuzzy c-means formula. If the cluster result doesn't accord with the structure of information, there should be one or a lot of clusters that are incorrectly separated leading to some clusters on the point of one another and error in cluster information. FCM enhancements to traditional c-means to handle such limitations and error information. Planned formula minimize error and increase accuracy of the perform cluster. PVCA s gets fine data in Hypothyroidism Dataset, E_coili_Dataset, Breastcancer_Dataset, yeast dataset. Fuzzy Clustering Algorithm is more average error rate as compare to proposed vector space model clustering algorithm (PVCA).

KEYWORDS: Data Mining, Clustering, K-means Algorithm, Fuzzy C-means algorithm, E.Colidataset, Yeast dataset.

I. INTRODUCTION

Clustering is that the method of distribution a homogenous cluster of objects into subsets known as clusters, so objects in every cluster are additional the same as 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 [1]. Cluster algorithms may be usually classified into 2 main categories, namely, supervised cluster and unsupervised cluster wherever the parameters of classifier are optimized. Several unsupervised cluster algorithms are developed. One such algorithmic rule is k-means, that assigns n objects to k clusters by minimizing the addition of square geometrician distance between the objects in every cluster to the cluster center. the most disadvantage of the k-means algorithmic rule is that the result's sensitive to the choice of initial cluster centroids and will converge to native optima [2].fast and sturdy cluster algorithms play a very important role in extracting helpful data in massive databases. The aim of cluster analysis is to partition a collection of N object into C clusters specified objects at intervals cluster ought to be the same as one another and objects in numerous clusters are ought to be dissimilar with every other[3]. Clustering is wont to quantize the out their information, to extract a group of cluster prototypes for the compact illustration of the dataset, into homogenized subsets. cluster may be a mathematical tool that tries to get structures or bound patterns in an exceedingly dataset, wherever the objects within every cluster show an explicit degree of similarity. It is achieved by varied algorithms that take issue considerably in their notion of what constitutes a cluster and the way to expeditiously realize them. Cluster analysis isn't an automatic task, however an unvaried method of data discovery or interactive multi-objective improvement. It'll usually necessary to switch preprocessing and parameter till the result achieves the specified properties. In cluster, one amongst the foremost wide used algorithms is fuzzy cluster algorithms. Fuzzy pure mathematics was 1st planned by Zadeh in 1965 & it gave a thought of uncertainty of happiness that was delineated by a membership perform. the employment of fuzzy set provides general category membership perform. Applications of fuzzy pure mathematics in cluster analysis were early planned within the work of bellhop, Zadeh, and Ruspini This paper opens door step of fuzzy cluster [4]

Clustering: agglomeration is a vital job in knowledge analysis and data processing applications. Information divides into similar object teams supported their options by agglomeration method. Every information cluster with similar objects is clusters. It means that clusters square measure the ordered set of information that has the acquainted characteristics. Agglomeration could be a method of unsupervised learning. Extremely superior clusters have high intra-class similarity and low inter-class similarity. Agglomeration algorithms have several classes like hierarchical- primarily based algorithms, partition-based algorithms, density-based algorithms and grid primarily based algorithms. Partition-based cluster: it's centroid based mostly clustering during which information points splits into k partition and every partition represents a cluster. Completely different ways of partitioning cluster are k-means, bisecting k-means technique and therefore the Probabilistic centroid and FCM.

II. TYPES OF CLUSTERING

KMC: K-means cluster technique could be a technique of cluster that is wide used. This formula is that the hottest cluster tool that's utilized in scientific and industrial applications. it's a technique of cluster analysis that aims to partition ‘n’ observations into k clusters during which every observation belongs to the cluster with the closest mean [5]

FCM: Fuzzy cluster could be a powerful unattended technique for the analysis of information and construction of models. In several things, fuzzy cluster is additional natural than hard cluster. Objects on the boundaries between many categories don't seem to be forced to completely belong to at least one of the categories, however rather are appointed membership degrees between zero and one indicating their partial membership. Fuzzy c-means formula is most generally used. Fuzzy c-means cluster was 1st according within the literature for a special case (m=2) by Joe Dunn in 1974. the overall case (for any m bigger than 1) was developed by Jim Bezdek in his Ph.D. thesis at university in 1973. It may be improved by Bezdek in 1981. The FCM employs fuzzy partitioning specified an information purpose will belong to all or any teams with completely different membership grades between zero and 1[6].

Applications of agglomeration

Fuzzy C-Means Grouping for knowledge division, agglomeration approach for cooperative filtering reference, agglomeration approach in Master Card fraud Detection, agglomeration technique for Diseases Analysis and medication, In Grouping of High-Dimensional records with Reference Similarity[7]

III. LITERATURE SURVEY

Bara’a A et al. [8] discovered that performance of clustering algorithms degrades with more and more overlaps among clusters in a data set. These facts have motivated to develop a fuzzy multi-objective particle swarm optimization framework (FMOPSO) in an innovative fashion for data clustering, which is able to deliver more effective results than state-of-the-art clustering algorithms. To ascertain the superiority of the proposed algorithm, number of statistical tests has been carried out on a variety of numerical and categorical real-life data sets.

D Małyszko et al.[9] proposed adaptive rough entropy clustering algorithms in image segmentation. Incorporating the most important image data information into the segmentation process has resulted in the development of innovative frameworks such as fuzzy systems, rough systems and recently rough - fuzzy systems. Rough entropy framework proposed in has been dedicated for application in clustering systems, especially for image segmentation systems.

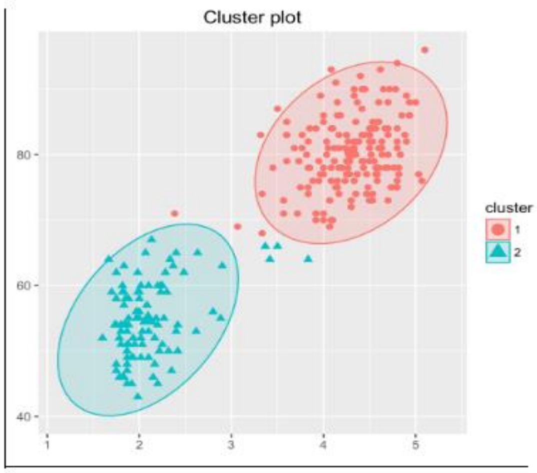

Fig 1 Partition-based clustering

Lingzi Duan et al. [10]. Sensitive to the initial number and centers of clusters is one shortcoming of fuzzy c-means clustering method. Aiming to reduce the sensitivity, a partial supervision-based fuzzy c-means clustering method is proposed in this paper. In this method, the data is first clustered with standard fuzzy c-means algorithm. If the clustering result doesn’t accord with the structure of data, there must be one or more 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. In both situations, approaches are then proposed in this new method to update the appropriate cluster number and cluster centers. With the updated cluster centers as labeled patterns, partially supervised fuzzy clustering is carried to give the appropriate clusters. Experiments on four 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.

S K Dubey et al. [11] present a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms. The factors use in this work upon which the

Received 3 October 2018, Revised Received, 9 November 2018, Accepted 19 November, 2018, Available online 25 November , 2018

behavior patterns of both the algorithms analyze are the numbers of data points as well as the number of clusters. The result of this comparative study is that FCM produces closer result to the K-means but still computation time is more than k-means due to involvement of the fuzzy measure calculations.

W.H. Au et al. [12] applied fuzzy linguistic terms to relational databases with numerical and categorical attributes. Later, they proposed the F-APACS method (1998) to discover fuzzy association rules. They utilized adjacent difference analysis and fuzziness in finding the minimum support and confidence values instead of having them supplied by a user. They determine both positive and negative associations.

T.P. Hong et al. [13] proposed an algorithm that integrates fuzzy set concepts and Apriori mining algorithm to find interesting fuzzy association rules from given transactional data. In another paper, proposed definitions for the support and confidence fuzzy membership grades and designed a data mining approach based on fuzzy sets to find association rules with linguistic terms of human knowledge.

Shital Shah et al. [14]. Cancer leads to approximately 25% of all mortalities, making it the second leading cause of death in the United States. Early and accurate detection of cancer is critical to the well-being of patients. Analysis of gene expression data leads to cancer identification and classification, which will facilitate proper treatment selection and drug development. Gene expression data sets for ovarian, prostate, and lung cancer were analyzed in this research. An integrated gene-search algorithm for genetic expression data analysis was proposed. This integrated algorithm involves a genetic algorithm and correlation-based heuristics for data preprocessing (on partitioned data sets) and data mining (decision tree and support vector machines algorithms) for making predictions. Knowledge derived by the proposed algorithm has high classification accuracy with the ability to identify the most significant genes. Bagging and stacking algorithms were applied to further enhance the classification accuracy. The results were compared with that reported in the literature. Mapping of genotype information to the phenotype parameters will ultimately reduce the cost and complexity of cancer detection and classification.

Kulkarni U. V. et al. [15] unsupervised fuzzy min-max clustering neural network in which clusters are implemented as fuzzy set using membership function with a hyper box core that is constructed from a min point and a max point. In the sequel to Min-Max fuzzy neural network classifier, proposed fuzzy hyper line segment clustering neural network (FHLSCNN). The FHLSCNN first creates hyper line segments by connecting adjacent patterns possibly falling in same cluster by using fuzzy membership criteria. Then clusters are formed by finding the centroids and bunching created HLSs that fall around the centroids.

IV SIMULATION ON MATLAB

The Performance analysis of MATLAB version (R2013a) i.e. used for this thesis Implementation of information mining provides processor optimized libraries for quick execution and computation and performed on input cancer dataset . It uses its JIT (just in time) compilation technology to supply execution speeds that rival traditional programming languages. It may also additional advantage of multi core and digital computer computers, MATLAB give several multi-threaded algebra and numerical operate. These functions automatically execute on multiple process thread during a single MATLAB, to execute quicker on multicore computers. During this thesis, all increased efficient information retrieve results were performed in MATLAB (R2013a). MATLAB is that the high-level language and interactive environment utilized by a lot of engineers and scientists worldwide. It lets the explore and visualize concepts and collaborate across totally different disciplines with signal and image process, communication and computation of results. MATLAB provides tools to accumulate, analyze, and visualize information, modify you to induce insight into your information during a division of the time it'd take exploitation spreadsheets or traditional programming languages. It may also document and share the results through plots and reports or as printed MATLAB code. MATLAB (matrix laboratory) could be a multi paradigm numerical computing scenario and fourth generation programming language. It’s developed by math work; MATLAB permits matrix strategy, plotting of operates and knowledge, implementation of rule, construction of user interfaces with programs. MATLAB is meant in the main for mathematical computing; an optional tool box uses the MuPAD symbolic engine, permitting access to symbolic computing capabilities. It’s simulating on mat research lab (R2013a) for this work we use Intel 1.4 GHz Machine and OS window7, window-xp etc. MATLAB version (R2013a) could be a high-level technical calculate language and interactive environment for rule development, knowledge visual image, records analysis, and numeric computation Mat research lab could be a software system program that permits you to try to information manipulation and visual image, calculations, math’s and programming. It may be accustomed do terribly easy still as very refined tasks. Database, analysis, visual image, and rule development. You’ll perform efficient information retrieve improvement. Many functions within the toolbox are

International Journal of Innovative Research in Technology & Science
Received 3 October 2018, Revised Received, 9 November 2018, Accepted 19 November, 2018, Available online 25 November , 2018
ISSN: 2321-1156 Volume VI Issue VI, October 2018

multithreaded to take benefit of multicore and multiprocessor computers.

V. RESULT ANALYSIS

(I). Performance comparison between FCMCA & PVCA using Breastcancer_Dataset

Table1 Performance comparison between FCMCA & PVCA

Breast Cancer Dataset			
Initial Set Random Values	**Technique**	**Time (in sec)**	**Error Rate (in %)**
0.68	FCMCA	7.47245	4.63243
	PVCA	10.9357	1.76781
0.167	FCMCA	8.65806	4.32226
	PVCA	6.56764	1.17052
0.1152	FCMCA	6.53644	3.81701
	PVCA	5.00763	1.04324
0.9	FCMCA	.561604	2.6552
	PVCA	1.74721	.0087105

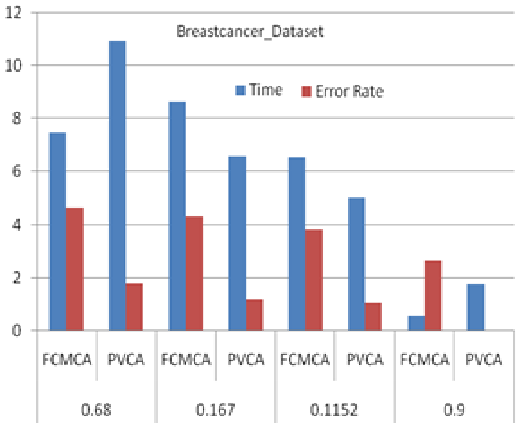

Fig 2 Performance analysis between FCMCA & PVCA

They have study freshly analysis paper within the field of knowledge mining and determine numerous challenge and objective to figure within the field of an improved and minimize error victimization planned cluster rule and FCM .An increase accuracy of the operate cluster new technique. A requirement to use caution as increasing n ends up in smaller error-function values by definition.

(II). Results Graph based on Breastcancer_Dataset

Result analysis based on breast cancer dataset used and initial set random values. FCMCA is more error rate as compare to PVCA .but FCMCA time take minim as compare to PVCA. PVCA is better as compare to FCMCA because data redundancy is more but PVCA is minim redundancy .it is show figure 2, below graph show time is more but error minimum. PVCA s gets fine data in breast cancer dataset.

(III) Performance comparison between FCMCA & PVCA using Hypothyroidism Dataset.

Table 2 Performance comparison between FCMCA & PVCA

Hypothyroidism Dataset			
Initial Set Random Values	**Technique**	**Time (in sec)**	**Error Rate (in %)**
0.86	FCMCA	6.92644	4.2227
	PVCA	11.2009	1.37214
0.18	FCMCA	5.55364	3.83872
	PVCA	2.87042	1.12319
0.52	FCMCA	5.46004	4.00499
	PVCA	4.75803	1.21675
0.07	FCMCA	8.12765	3.76968
	PVCA	7.59725	1.12319

(iv) Results Graph based on Hypothyroidism Dataset

Result analysis based on hypothyroidism dataset used and initial set random values. FCMCA is more error rate as compare to PVCA .but FCMCA time take minim as compare to PVCA. PVCA is better as compare to FCMCA because data redundancy is more but PVCA is minim redundancy .it is show figure 3, below graph show but error minimum. PVCA gets fine data in hypothyroidism dataset.

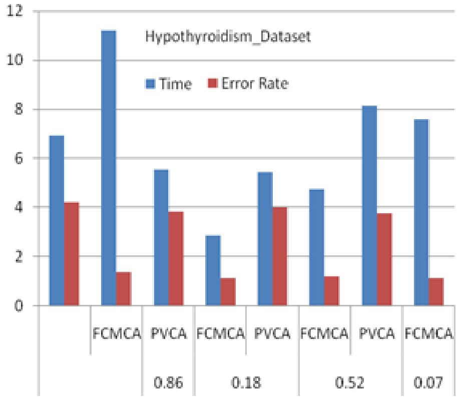

Fig3 Performance analysis between FCMCA & PVCA

VI. Conclusion

Our analysis work needs find in result optimum answer and given fully totally different cluster .In medical data analysis supported our fuzzy clustering formula is additional average error rate as compare to proposed vector space model clustering formula (PVCA) in given completely different grouping methodology is applied on dataset record. FCM rule can be a particular cluster rule, has been exploited in intensive vary of

Received 3 October 2018, Revised Received, 9 November 2018, Accepted 19 November, 2018, Available online 25 November , 2018
ISSN: 2321-1156 Volume VI Issue VI, October 2018

engineering and scientific disciplines, as an example, medicine imaging, pattern detection, processing and bioinformatics. Visible of the particular truth, the initially developed FCM makes use of the squared-norm to figure out the similarity between prototypes and data points, and it performs well only at intervals the case of cluster spherical clusters. Several algorithms are developed by numerous authors supported the FCM however error additional with the aim of cluster minimize general dataset. Proposed technique recognizes some vital data points discover in dataset and improvement. Cluster technique to achieve a great deal of accuracy at intervals the result and reduce the time taken for data or data retrieval from large dataset. Proposed formula is improved clustering as compare to fuzzy c-mean formula as a result of additional accuracy supported minim error rate show in result. it's known as higher original centroid using PVCA. The performance analysis best technique additionally a proposed vector space model clustering formula (PVCA) has been compared and analysis with existing algorithms but this has well-tried to be additional efficient in terms of quality. The results made were satisfactory in terms of very well accuracy.

REFERENCES

[1]. M. Yuwono, S. W. Su, B. D. Moulton, and H. T. Nguyen, “Data clustering using variants of rapid centroid estimation,” IEEE Transactions on Evolutionary Computation, vol. 18, no. 3, pp. 366–377, 2014.

[2]. S. C. Satapathy, G. Pradhan, S. Pattnaik, J. V. R. Murthy, and P. V. G. D. P. Reddy, “Performance comparisons of PSO based clustering,” Inter JRI Computer Science and Networking, vol. 1, no. 1, pp. 18–23, 2009.

[3]. M. S. Yang,” A Survey of fuzzy clustering” Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.

[4]. A. vathy-Fogarassy, B. Feil, J. Abonyi “Minimal Spanning Tree based Fuzzy clustering” Proceedings of World academy of Sc., Eng & Technology, vol8, Oct-2005, 7-12.

[5]. Amandeep Kaur Mann, Navneet Kaur Mann, ìReview Paper On Clustering Techniquesî ,Global Journal Of Computer Science And Technology Software & Data Engineering, VOL. 13 ,2013

[6]. Pal N.R, Pal K, Keller J.M. and Bezdek J.C, “A Possibilistic Fuzzy c-Means Clustering Algorithm”, IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

[7]. Kamran Shaukat Dar, Imran Javed, Warda Amjad, Aroosa Shamim, “Survey of Clustering Applications”, Journal of Network Communications and Emerging Technologies (JNCET), Volume 4, Issue 3, October ,2015.

[8]. Bara’a Ali Attea, “A fuzzy multi-objective particle swarm optimization for effective data clustering,” Springer-July 2010, pp. 305-312.

[9]. Dariusz Małyszko, Jarosław Stepaniuk “Adaptive Rough Entropy Clustering Algorithms in Image Segmentation”, pp. 199-231,2010.

[10]. Lingzi Duan, Fusheng Yu, Li Zhan, “An Improved Fuzzy C-means Clustering Algorithm”, 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD.

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

[12]. W.H. Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, in: Proc. IEEE Internat. Conf. Fuzzy Systems, 1998, pp. 1314–1319.

[13]. T.P. Hong, C.S. Kuo, S.C. Chi, A fuzzy data mining algorithm for quantitative values, in: Proc. Internat. Conf. Knowledge-Based Intelligent Information Engineering Systems, 1999, pp. 480–483.

[14]. Shital Shah, Andrew Kusiak, “Cancer gene search with data-mining and genetic algorithms”, Computers in Biology and Medicine 37 , 251 – 261,2007.

[15]. U. V. Kulkarni, T. R. Sontakke, and A. B. Kulkarni, “Fuzzy hyperline segment clustering neural network”, Electronics Letters, IEEE, vol. 37, no. 5, pp. 301–303, March. 2001.

[16]. Lior Rokach and Oded Maimon,“Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)”, ISBN: 981-2771-719, World Scientific Publishing Company, , 2008.

[17]. Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, “The Applications of Genetic Algorithms in Stock Market Data Mining Optimization”, Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications, pp- 593-604,sept 2005

[18]. M. Craven and J. Shavlik, “Learning rules using ANN “, Proceeding of 10th International Conference on Machine Learning, pp.-73-80, July 1993.

[19]. S. Nithya, G. Shine Let, “ Bio-Medical Image Retrieval Using SVM,” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 10, December 2012.

[20]. Khaled Hammouda. A Comparative Study of Data Clustering Techniques. www.pami.uwaterloo.ca /pub/ hammouda/sde625-paper.pdf , 1-21.

[21]. Kamran Shaukat Dar, Imran Javed, Warda Amjad, Aroosa Shamim, “Survey of Clustering Applications”, Journal of Network Communications and Emerging Technologies (JNCET), Volume 4, Issue 3, October ,2015

[22]. Bara’a Ali Attea, “A fuzzy multi-objective particle swarm optimization for effective data clustering,” Springer-July 2010, pp. 305-312.

Received 3 October 2018, Revised Received, 9 November 2018, Accepted 19 November, 2018, Available online 25 November , 2018

[23].Dariusz Małyszko, Jarosław Stepaniuk “Adaptive Rough Entropy Clustering Algorithms in Image Segmentation”, pp. 199-231,2010.

[24].Lingzi Duan, Fusheng Yu, Li Zhan, “An Improved Fuzzy C-means Clustering Algorithm”, 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD.

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

[26].W.H. Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, in: Proc. IEEE Internat. Conf. Fuzzy Systems, 1998, pp. 1314–1319.

[27].T.P. Hong, C.S. Kuo, S.C. Chi, A fuzzy data mining algorithm for quantitative values, in: Proc. Internat. Conf. Knowledge-Based Intelligent Information Engineering Systems, 1999, pp. 480–483.