International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume3, Issue 1 (January'2017)

A REVIEW OF FUZZY CLUSTERING ON BIG DATA

M.M.Kavitha,
Research Scholar,
Bharathiar University,
Coimbatore, Tamilnadu,India.

Dr. B.Anandhi,
Associate Professor and Head,
Department of Computer Science,
Vellalar College for Women,
Erode,Tamilnadu,India.

Abstract: With the exponential growth of data from various social networks like Facebook, Twitter, Mobile applications, Digital cameras, Sensor networks etc., and also from biomedical researches the overall data volume has increased tremendously. So analyzing and extracting fruitful information from such a dynamic data is very much challenging task today. Data grouping or clustering plays a vital role in handling big data which is the basic foot step in data mining, pattern recognition and also in medical predictions. The clustering techniques are very much suitable for handling big data in this case the learning parameters are computed from learning data. Clustering approaches can be classified into two categories namely- Hard clustering and soft clustering. In hard clustering data is divided into clusters in such a way that each data item belongs to a single cluster only while soft clustering also known as fuzzy clustering forms clusters such that data elements can belong to more than one cluster based on their membership levels which indicate the degree to which the data elements belong to the different clusters. This paper deals with an attempt at studying the data clustering algorithms based on fuzzy techniques. These fuzzy clustering algorithms have been widely studied and applied in a variety of substantive areas.

Keywords: Big Data, Fuzzy Clustering, C-Means, K-means, KNN

I.INTRODUCTION

Big data from various resources are need to be collected and analyzed for decision making based on the needs of user. The data has 3V's such as Volumes, Velocity and Variety. The big volume of data needs effective handling techniques for managing and reusing of data based on the analytical aspects. This massive volume of data can be really useful. But they are very much problematic in terms of their storage and analysis. The big volume makes analytical operations, process operations, retrieval operations very difficult and they are very much time consumable. In order to overcome these problems we need to have clustered form of big data [1].The fuzzy based clustering technique will improve the accuracy of the clustering process. Clustering is a foremost responsibility notion of data mining and it is task of separating piece of information points into homogeneous module so that items is similar means those points put into same class and different means those points put into different classes.

The term "similarity" should be understood as mathematical similarity, calculated in a number of well-defined intelligence. In metric spaces, similarity is often defined by means of a distance. Clustering is an unsupervised [2] learning which is finding groups from unlabeled data and the class-prediction is done on unlabeled facts after a supervised learning on pre-labeled data. Clustering can find different types of similarity measures may be used to identify classes depending on the data and its application, in clustering, information is divided into crisp clusters and each data belongs to exactly one cluster. Hard clustering methods are based on classical set theory, and require that an object either does or does not belong to a cluster. It is a separation of data into a specified number of mutually exclusive subsets. The four main classes of clustering algorithms available in the literature are partitioning methods, hierarchical methods, density-based clustering and grid-based clustering. Clustering adds to the value of living databases by helpful secreted relationships in the data. Clustering also called as the hard clustering, data is divided into distinct clusters.

Soft computing differs from conventional computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth, and approximation. In effect, the role model for soft computing is the human mind. Fuzzy systems are appropriate for vague or fairly accurate reasoning, particularly for the system with a mathematical model that is difficult to derive. Fuzzy logic allows decision making with predictable values under vague information. Fuzzy Logic is a compilation of notebooks and packages and that are planned to bring in fuzzy set theory with fuzzy logic in the mathematical area and which is provides a powerful tool for studying fuzzy logic and developing fuzzy applications in many area. In fuzzy clustering, the [3] data points can belong to multiple clusters, and associated with each of the elements are membership marks which point out the degree to which the data points fit in to the unlike clusters. Fuzzy clustering is now a mature and vibrant area of research with highly innovative advanced applications. Encapsulating this through presenting a careful selection of research contributions, this book addresses timely and relevant concepts and methods, whilst identifying major challenges and recent developments in the area.

Fuzzy clustering methods [4] allow the objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The discrete nature of the hard partitioning also causes difficulties with algorithms based on analytic functional, since these functional are not differentiable. Fuzzy clustering is a progression of conveying these membership levels, and then [5] using them to assign data elements to one or more clusters.

II LITERATURE REVIEW

RuiqiongCai et al., [6] focuses on the clustering of temporal dataset, and presents a new version of constraints-equipped fuzzy C-Means algorithm, where the temporal constraints are described by fuzzy sets rather than crisp intervals. This new version fuzzy c-means overcomes the disadvantages of the old version where strict requirements on the choosing of the parameters are needed and in many cases the result is not so satisfactory.

Celikyilmaz et al., [7] proposed a new fuzzy system modeling approach based on improved fuzzy functions to model systems with continuous output variable. The new modeling approach introduces three features: i) an Improved Fuzzy Clustering (IFC) algorithm, ii) a new structure identification algorithm, and iii) a nonparametric inference engine. The IFC algorithm yields simultaneous estimates of parameters of c-regression models, together with fuzzy cpartitioning of the data, to calculate improved membership values with a new membership function. The structure identification of the new approach utilizes IFC, instead of standard fuzzy C-Means clustering algorithm, to fuzzy partition the data, and it uses improved membership values as additional input variables along with the original scalar input variables for two different choices of regression methods: least squares estimation or support vector regression, to determine ldquofuzzyfunctionsrdquo for each cluster. With novel IFC, one could learn the system behaviour more accurately compared to other FSM models. The nonparametric inference engine is a new approach, which uses the alike -nearest neighbour method for reasoning.

Jiabin Deng et al., [8] proposed an improved fuzzy clustering-text clustering method based on the fuzzy CMeans clustering algorithm and the edit distance algorithm. The author used the feature evaluation to reduce the dimensionality of high-dimensional text vector. Because the clustering results of the traditional fuzzy C-Means clustering algorithm lack the stability, the author introduced the highpower sample point set, the field radius and weight. Due to the boundary value attribution of the traditional fuzzy CMeans clustering algorithm, the author recommended the edit distance algorithm.

Sato-Ilic et al., [9] proposed a generalized kernel fuzzy clustering model and investigated the features of the proposed model. An additive clustering model has been proposed that considers the overlapping of clusters whose target data is similarity data. In addition, by introducing the concept of a fuzzy cluster to the additive clustering model, an additive fuzzy clustering model has been proposed. In these models, sharing common properties of clusters combine rdquoadditivelyrdquo and the given similarity between a pair of objects is estimated as the sum of the shared common properties. Therefore, in these models, the effects of the interaction of different clusters cannot be considered. In order to solve this problem, the author proposed a generalized kernel fuzzy clustering model which is an extension of the additive fuzzy clustering model to a nonlinear fuzzy clustering model through the use of kernel functions. In this new model, the degree of objects to clusters is estimated in a mapped higher dimensional space using kernel functions.

Chatzis et al., [10] proposed a novel FCM-type fuzzy clustering scheme providing two significant benefits when compared with the existing approaches. First, it provides a well-established observation space dimensionality reduction framework for fuzzy clustering algorithms based on factor analysis, allowing concurrent performance of fuzzy clustering and, within each cluster, local dimensionality reduction. Secondly, it exploits the outer tolerance advantages of SMMs to provide a novel, soundly founded, nonheuristic, robust fuzzy clustering framework by introducing the effective means to incorporate the explicit assumption about student's t-distributed data into the fuzzy clustering procedure.

Jian Yu et al., [11] proposed a Generalized Fuzzy Clustering Regularization (GFCR) model and then studied its theoretical properties. GFCR unifies several fuzzy clustering algorithms, such as Fuzzy C-Means (FCM), Maximum Entropy Clustering (MEC), fuzzy clustering based on Fermi-Dirac entropy, and fuzzy bidirectional associative clustering network, etc. The proposed GFCR becomes an alternative model of the Generalized FCM (GFCM) that was recently proposed by Yu and Yang. To advance theoretical study, the authors have the three following considerations. 1) The author gave an optimality test to monitor if GFCR converges to a local minimum. 2) The author related the GFCR optimality tests to Occam's razor principle, and then analyzed the model complexity for fuzzy clustering algorithms. 3) The author offered a general theoretical method to evaluate the performance of fuzzy clustering algorithms.

Yanfeng Zhang et al., [12] presented a new Neighbour Sharing Selection based Agglomerative fuzzy K-Means (NSS-AKmeans) algorithm for learning optimal number of clusters and generating better clustering results. The NSSAKmeans can identify high density areas and determine initial cluster centers from these areas with a neighbor sharing selection method. To select initial cluster centers, the author proposed Agglomeration Energy (AE) factor for representing global density relationship of objects, and a Neighbours Sharing Factor (NSF) for estimating local neighbour sharing relationship of objects. Then the author used the Agglomerative Fuzzy K-Means clustering algorithm

to further merge these initial centers further to obtain the preferred number of clusters and generate better clustering results.

Yang et al., [13] proposed an efficient data clustering algorithm. It is well known that K-Means (KM) algorithm is one of the most popular clustering techniques because it is unproblematic to implement and work rapidly in most situations. But the sensitivity of KM algorithm to initialization makes it effortlessly trapped in local optima. KHarmonic Means (KHM) clustering resolves the problem of initialization faced by KM algorithm. Even then KHM also easily runs into local optima. PSO algorithm is a global optimization technique. A hybrid data clustering algorithm based on the PSO and KHM (PSOKHM) was proposed by Yang .This hybrid data clustering algorithm utilizes the advantages of both the algorithms. Therefore the PSOKHM algorithm not only helps the KHM clustering run off from local optima but also conquer the inadequacy of the slow convergence speed of the PSO algorithm.

Zadeh et al., [14] proposed type-2 fuzzy logic in 1975 .It is a generalization of type-1 fuzzy set and consist in using uncertain (or fuzzy) membership function. Unlike fuzzy logic, which uses 2-dimensional memberships, type-2 fuzzy logic is 3-dimensional memberships, so it can handle more uncertainties and it is better for identifying outliers. For type-2 fuzzy logic the membership function is called FOU (Footprint of Uncertainty) and it is bounded by two membership function (upper membership function and lower membership function). There are two categories: generalized type-2 fuzzy uses a 3-dimensional FOU and interval type-2 fuzzy using a 2-dimensional FOU. The latter is the most used due to its simplicity and reduced complexity.

Keller et al., [15] where class memberships are given to the sample, as a function of the sample's distance from its K Nearest Neighboring training samples. A Fuzzy K-NN Classifier is one of the most successful techniques for applications due to its simplicity and also because of giving some information about the certainty of the classification decision. Keller et al assume that the improvement of the error rate might not be the major advantage from using the FKNN model. More importantly, the model offers a percentage of certainty which can be used with a "refuse-to-decide" option. Thus objects with overlapping classes can be detected and processed individually.

Lu,W et al., [16] The k-Nearest Neighbor method is placed in the top ten data mining techniques. Prajesh P, Anchalia, and Kaushik Roy used this well know classification technique to classify big data. They run this method on an apache Hadoop environment that of course uses the Map Reduce paradigm to process on big data. They faced a lot of problems and the most important one is balancing between the friendly user interface and performance. They implemented the k nearest neighbor on the Hadoop using multiple computers to delete the limitations of computational capability and speeding up the processing time by having groups of systems working together and connected over a network. They also compared their results using a MapReduce K nearest neighbor with sequential K nearest neighbor and concluded that the map reduce k nearest neighbor gives better performance than the sequential K nearest neighbor with big data.

Jain and Dubes et al., [17] argue the typical pattern clustering activity involves pattern representation (optionally including feature extraction and/or selection), the definition of a pattern proximity measure appropriate to the data domain, clustering or grouping, data abstraction (if needed), and assessment of output (if needed). The approach to clustering in vogue broadly fall under statistical, fuzzy and machine learning techniques. The statistical techniques analyses the data's linear characteristics and classifies it accordingly. The fuzzy set theory technique introduces uncertainty similar to human thinking in the classification process and thus making it robust. The machine learning technique such as artificial neural network (ANN) captures the nonlinear characteristics of data, resulting in better classification.

Bezdek et al., [18] propounded fuzzy c-means (FCM) algorithm, which is one of the best known fuzzy clustering approaches. Its use for various applications is well described and analyzed. FCM method works on the optimization of a specific cost function, and it operates well when the clusters are compact or isotropic. Various variants of FCM have emerged as some researchers have worked to decrease the time of consumption others have worked to improve the accuracy. This method has been applied to various data types particularly in the area of image segmentation with slight variations.

Nasullah Khalid Alham et al., [19] used a MapReduce support vector machine technique. They named this technique MRSMO technique (MapReduce based distributed SVM algorithm for automatic image annotation). Their technique depends on partitioning the training data into subsets and sent these subsets across groups of computers. They evaluated their technique in an experimental environment and it result in a significant reduction in the training time and a high accuracy in both binary and multiclass classification.

Ke Xu et al., [20] used the parallel Support Vector Machine based on MapReduce method for classification of emails which is a big data set. They implement an experiment on this data set and used many techniques in evaluation but the support vector machine based on MapReduces show a significant reduction in the training time. Big data sets are very complex to be analyzed using classical Support Vector Machine but the parallel Support Vector Machine depending on MapReduce and can deal easily with big data. The MapReduce distribute the subsets of the training data among many nodes to improve the computation speed and improve the accuracy.

Gayatri Nair et al., [21] The proposed system has been designed using one of the most popular approaches for big

data; MapReduce framework, which is a functional programming paradigm that is well suited to handle parallel processing of huge data sets. The decision tree technique in FID3 algorithm involves constructing a tree to model the classification process. Thus, understandable prediction rules are created from the training data, which helps in classification process more easily. The output results showed that using this system the admin can get an idea about the students enrolment in next level of education. From the listed universities for a student, he/she can choose any of the university and go on with further admission process. Validation of multiple conditions and criteria improves the performance of the prediction of universities for the students according to their scores and university cut-offs. Thus, the concept of predictive analytics holds great promise for helping educational institution make evidence- based decisions related to students.

Malak El Bakry et al., [22] The proposed algorithm consists of two parts; the mapper and the reducer. The mapper algorithm is used to divide the data sets in to chunks over the computing nodes and produce a set of intermediate records. These records produced by the map function take the form of a "(key, data)" pair. Mapper in the individual nodes execute the computing process and send the results to the reduce function. The reducer algorithm receives the results of individual computations and put them together to obtain the final result. Good accuracy of the performance was obtained using the Fuzzy K-Nearest Neighbor method. Future work will concentrate on enhancing the results using Fuzzy techniques in the reducer rather than using the mode.

III.CONCLUSION

Clustering techniques play a key role in many applications. Many researches are being done in this area for the betterment of the overall performance of the clustering techniques. Clustering is a potential technique in many data mining applications. This survey concentrates on efficiency of the clustering approaches. This survey utilized a clustering algorithm to provide the best clustering results with greater clustering accuracy and reduced mean squared error and execution time, respectively with quick convergence. As the survey is done on the data clustering based on fuzzy techniques hence it is the most efficient technique when compared with the clustering techniques.

IV.REFERENCES

[1]. K. Velusamy and R.Manavalan , "Performance Analysis of Unsupervised Classification based on Optimization," International Journal of Computer Applications (0975 – 8887, Volume 42– No.19, March 2012.

[2]. Raju. G,Thomas.B, Tobgay. S and Kumar T.S, "Fuzzy Clustering Methods in Data Mining: A Comparative Case Anaylisis," International Conference on Advanced Computer Theory and Engineering, 2008. ICACTE '08 (489-493),ISBN : 978-0-7695- 3489-3.

[3]. Dilip kumar praihar and sanjib Chandra de sharker , "A comparative study of fuzzy c means algorithm and entropy based fuzzy clustering algorithms," computing and informatics, vol 30,2011, 701-720.

[4]. S.P Priyadharshini and Ramachandra V. Pujeri, "Performance of Fuzzy Clustering based Optimization ," International Journal of Advances in Engineering & Technology, May 2014, ISSN 22311963.

[5]. Jianxiong Yang and Junzo Watada, "fuzzy clustering analysis of datamining : application of an accident mining system ," International journal of innovative computing, information and control icic international 2012-ISSN 1349-4198.

[6]. RuiqiongCai and Fusheng Yu, "Fuzzy Temporal Constraints Based Fuzzy Clustering Algorithm for Temporal Dadaset", Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '09), Vol. 1, Pp. 480–484, 2009.

[7]. Celikyilmaz A and BurhanTurksen I, "Enhanced Fuzzy System Models With Improved Fuzzy Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 16, No. 3, Pp. 779–794, 2008.

[8]. Jiabin Deng, JuanLi Hu, Hehua Chi and Juebo Wu, "An Improved Fuzzy Clustering Method for Text Mining", Second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), Vol. 1, Pp. 65–69, 2010.

[9]. Sato-Ilic M, Ito S and Takahashi S, "Generalized kernel fuzzy clustering model", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Pp. 421–426, 2009.

[10]. Chatzis S and Varvarigou T, "Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using t-Distributions", IEEE Transactions on Fuzzy Systems, Vol. 17, No. 3, Pp. 505–517, 2009.

[11]. Jian Yu and Miin-Shen Yang, "A Generalized Fuzzy Clustering Regularization Model With Optimality Tests and Model Complexity Analysis", IEEE Transactions on Fuzzy Systems, Vol. 15, No. 5, Pp. 904–915, 2007.

[12]. Yanfeng Zhang, XiaofeiXu and Yunming Ye, "NSSAKmeans: An Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number", 2nd International Conference on Advanced Computer Control (ICACC), Vol. 2, Pp. 32–38, 2010.

[13]. Yinghua Zhou, Hong Yu and XuemeiCai, "A Novel k-Means Algorithm for Clustering and Outlier Detection", Second International Conference on Future Information Technology and Management Engineering (FITME '09), Pp. 476–480, 2009.

[14]. L. A. Zadeh, "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning–1," Information Sciences, vol. 8, pp. 199–249, 1975.

[15]. Keller, J.M., Gray, M.R., and Given, J.A., (1985). A Fuzzy K-Nearest Neighbor Algorithm. IEEE Trans. Syst., Man, Cybern., Syst., 15 (4), 580-585.

[16]. Lu, W., et al., Efficient processing of k nearest neighbor joins using MapReduce. Proceedings of the VLDB Endowment, 2012. 5(10): p. 1016-1027.

[17]. Jain, A. K. and Dubes, R. C. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ. 1988.
[18]. J. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. thesis, Ithaca, NY: Cornell University, 1974.
[19]. Nasullah Khalid Alham, Maozhen Li, Yang Liu, and Suhel Hammoud, (2011). a MapReduce-based distributed SVM algorithm for automatic image annotation
[20]. Xu, K., Wen, C., Yuan, Q., He, X., & Tie, J. (2014). A MapReduce based Parallel SVM for Email Classification. Journal of Networks JNW.
[21]. Gayatri Nair1, Shankar M. Patil2 Fuzzy Rule Based Classifier for Student Data using Map Reduce , International Journal of Advanced Research in Computer and Communication Engineering , Vol. 6, Issue 2, February 2017.
[22]. Malak El Bakry, Soha Safwat , Osman Hegazy, Big Data Classification using Fuzzy K-Nearest Neighbor, International Journal of Computer Applications (0975 – 8887) Volume 132 – No.10, December2015

