A Survey on Data Clustering Algorithms based on Fuzzy Techniques

L. Divya Sivanandini[1], M. Mohan Raj[2]

[1]Research Scholar, Department of Computer Science,
Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, India
dhivyan876@gmail.com

[2]Assistant Professor, Department of Computer Applications,
Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, India
mohanrajsns@gmail.com

Abstract: ***Clustering is used to describe methods for grouping of unlabeled data. Clustering is an important task in data mining to group data into meaningful subsets to retrieve information from a given dataset. Clustering is also known as unsupervised learning since the data objects are pointed to a collection of clusters which can be interpreted as classes additionally. The chief objective of the clustering is to present a collection of similar records. The clustering problem has been focused by many researchers. Data Clustering is a technique in which logically similar information is physically stored together. The numbers of disk accesses are to be minimized in order to increase the efficiency in the database systems. The basic data clustering problem might be defined as finding out groups in data or grouping related objects together. Many different clustering techniques have been proposed over the years such as Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods and Model-based methods. 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: Clustering, Data clustering, Fuzzy clustering.

1. Introduction

Due to the remarkable development of the modern information systems, the quantity of data that is collected has increased massively. In order to examine these enormous collections of data, the interdisciplinary field of Knowledge Discovery in Databases (KDD) has emerged. The fundamental step of KDD is called Data Mining [1]. Data Mining utilizes effective techniques to obtain interesting patterns and regularities from the data. Clustering is one of the most significant data mining techniques.

Clustering is the process of assembling the data records into significant subclasses (clusters) in a way that increases the relationship within clusters and reduces the similarity among two different clusters [2]. Other names for clustering are unsupervised learning (machine learning) and segmentation. Clustering is used to get an overview over a given data set. A set of clusters is often enough to get insight into the data distribution within a data set. Another important use of clustering algorithms is the preprocessing for some other data mining algorithm.

Cluster analysis is a most important technique for categorizing a 'mountain' of information into controllable meaningful piles. Cluster analysis is a data reduction tool that generates subgroups that are further controllable than individual datum. Similar to factor analysis, it observes the complete complement of inter-associations among variables. In cluster analysis there is no previous knowledge about which elements corresponding to each cluster. The grouping or clusters are defined through an analysis of the data. Subsequent multivariate analyses can be performed on the clusters as groups [3]

The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources. Clustering is a technique in which the objects are grouped by some means these objects are similar in characteristics. The chief objective of the clustering is to present a collection of similar records. In general, clustering and classification are often confused based on their functions, however there exists a huge difference between the clustering and classification. In case of classification, the objects are allocated to pre-defined classes, but in the case of clustering the classes are generated. The term "class" is commonly used as synonym to the term "cluster".

Cluster Analysis (CA) is an exploratory data analysis technique for managing collected data into significant taxonomies, groups, or clusters, according to the combinations, which increases the similarity of cases inside a cluster at the same time increasing the dissimilarity between the other groups that are primarily unknown. In this manner, CA generates new groupings without any fixed notion of what clusters might arise, while discriminant analysis categorizes people and items into previously known groups. CA gives no details regarding the existence of a particular cluster nor any analysis made. Thereby each cluster thus illustrates, regarding the data being collected, the class to which its corresponding members. Items in each cluster are the same based on certain parameters and dissimilar to those in other clusters.

Clustering is actually the categorization of objects into several groups, or more specifically, the partitioning of a data set into a subset of objects, with the intention that the data present in each subset possibly share certain similar features frequently proximity based on certain defined distance measure. Data clustering is a general approach for statistical data analysis, which is most commonly used in several ranges of areas, including bioinformatics, pattern recognition, image analysis and pattern recognition.[4]

Moreover the concept of clustering approach has been used in several fields like statistics, machine learning , data mining , operations research , medical diagnostics, facility location, and across multiple application areas including genetics, taxonomy, medicine, marketing, finance and e-commerce [5].The basic steps involved in clustering method are preprocessing and feature selection, similarity measure, clustering algorithm, result validation, result interpretation and application. Many different clustering techniques have been proposed over the years such as Partitioning clustering algorithms, Hierarchical methods, Density-based clustering algorithms, Grid-based methods, Model-based clustering algorithms. This paper carryout various types of data clustering algorithms which are developed by the researchers.

2. Objective of the Research

The primary aim of this research is to propose the novel technique for the clustering. The other goals of this research include the following:

- To develop a clustering technique that is not sensitive to initial positions of cluster centers.
- Developing a novel technique for clustering by using effective approaches which provide very high classification accuracy.
- To develop a clustering technique with very less error rate and with minimized execution time
- Developing a clustering technique with less convergence time and also with less number of iterations.
- To develop an effective clustering technique which provides better results in incomplete/noisy data.

3. Background Study

In order to promote efficient and effective clustering algorithm, substantial research work has been carried out in the recent years. A number of clustering schemes have been proposed by various authors to address the problems.

Chandra et al., [6] defined Clustering as a significant data mining function as it plays a key role in information retrieval in database operations. The SDBMS is characterized with some unique features and so has specific requirements in a clustering algorithm. Specifically the geographic databases in the SDBMS have noisy data and outliers, which the algorithm should handle effectively. Partitioning methods, Hierarchical methods and Density-based methods are the methodologies studied in this paper. The basic K-means, K-means with scalable framework and Continuous Kmeans are the Partitioning methods. CURE, BIRCH, and CHAMELEON are the Hierarchical clustering algorithms. DBSCAN and OPTICS are density based clustering algorithms. A Hypergraph based clustering algorithm is also studied that is an enhancement to the basic Density-based algorithm and the hypergraph itself is derived from the Delaunay triangulation of the spatial datasets. A recent clustering algorithm group's datasets into clusters with the spatial object as such as the core object and it utilizes Euclidean distance for that purpose. The algorithms are studied with respect to key factors such as dimensionality of data, worst case time, etc.

RuiXu [7] surveyed clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Several tightly related topics, proximity measure, and cluster validation, are also discussed in his work. Jain et al., [7] presented an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. They presented taxonomy of clustering techniques, and identified cross-cutting themes and recent advances. They also described some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

Parimala et al., [9] described a detailed survey of the existing density based algorithms namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN and DBCLASD based on the essential parameters needed for a good clustering algorithm. They analyzed the algorithms in terms of the parameters essential for creating meaningful clusters.

OrnellaCominetti et al.,[19] showed that the fuzzy spectral clustering method DifFUZZY performs well in a number of data sets, with sizes ranging from tens to hundreds of data points of dimensions as high as hundreds. This includes microarray data, where a typical size of a data set is dozens or hundreds (number of samples, conditions, or patients in medical applications) and dimension is hundreds or thousands (number of genes on the chip). The clustering methodology used in their approach isspecifically designed to handle non-Euclidean data sets associated with a manifold structure, asit seamlessly integrates spectral clustering approaches with the evaluation of cluster membershipfunctions in a fuzzy clustering context.

Yang [10] presented a survey of fuzzy clustering in three categories. The first category is the fuzzy clustering based on fuzzy relation. The second one is the fuzzy clustering based on objective function. Finally, we give 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.

Anil K. Jain [11] provided a brief overview of clustering, summarize well known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering. Clustering is in the eye of the beholder, so indeed data clustering must involve the user or application needs.

Yang et al., [12] 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. K-Harmonic 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 et al. in [13]. 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.

Clustering is a technique in data mining to find interesting patterns in a given dataset. A large dataset is grouped into clusters of smaller sets of similar data using K-Means algorithm. Initial centroids are required as input parameters when using K-Means clustering algorithm. There are different methods to choose initial centroids, from actual sample datapoints of a dataset. These methods are often implemented through intelligent agents, as the later are very commonly used in distributed networks given that they are not cumbersome for the network traffic. Moreover, they overcome network latency, operate in heterogeneous environment and possess fault-tolerant behaviour. A Multi Agent System (MAS) is proposed by Khan et al., [14] in this research work for the generation of initial centroids using actual sample datapoints. This multiagent system comprises four agents of K-Means clustering algorithm using different methods namely Range, Random number, Outlier and Inlier for the generation of initial centroids.

Chen Zhang et al., [15] presented a new clustering method based on K-Means that have avoided alternative randomness of initial center. This approach focused on K-Means algorithm to the initial value of the dependence of K selected from the aspects of the algorithm is improved. First, the initial clustering number is radicN. Second, through the application of the sub-merger strategy the categories were combined. The algorithm does not require the user is given in advance the number of cluster.

Clustering has taken its roots from algorithms like K-Means and K-Medoids. However conventional K-Medoids clustering algorithm suffers from many limitations. Firstly, it needs to have prior knowledge about the number of cluster parameter k. Secondly, it also initially needs to make random selection of k representative objects and if these initial k medoids are not selected properly then natural cluster may not be obtained. Thirdly, it is also sensitive to the order of input dataset. First limitation was removed by using cluster validity index. Aiming at the second and third limitations of conventional k-medoids, Pardeshi et al., [16] have proposed an improved k-medoids algorithm. In this work instead of random selection of initial k objects as medoids the author have proposed a new technique for the initial representative object selection. The approach is based on density of objects. The author finds out set of objects which are densely populated and choose medoids from each of this obtained set. These k data objects selected as initial medoids are further used in clustering process.

Yanfeng Zhang et al., [17] 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 NSS-AKmeans can identify high density areas and determine initial cluster centers from these areas with a neighbour 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.

Five clustering algorithms taken from the literature are reviewed, assessed and compared on the basis of the selected properties of interest by Baraldi et al., [18]. These clustering models are (1) Self-Organizing Map (SOM); (2) Fuzzy Learning Vector Quantization (FLVQ); (3) Fuzzy Adaptive Resonance Theory (fuzzy ART); (4) Growing Neural Gas (GNG); (5) Fully Self-Organizing Simplified Adaptive Resonance Theory (FOSART). Although our theoretical comparison is fairly simple, it yields observations that may appear parodoxical. First, only FLVQ, fuzzy ART, and FOSART exploit concepts derived from fuzzy set theory (e.g., relative and/or absolute fuzzy membership functions). Secondly, only SOM, FLVQ, GNG, and FOSART employ soft competitive learning mechanisms, which are affected by asymptotic misbehaviours in the case of FLVQ, i.e., only SOM, GNG, and FOSART are considered effective fuzzy clustering algorithms.

Jian Yu et al., [19] 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.

Factor analysis is a latent subspace model commonly used for local dimensionality reduction tasks. FCM type fuzzy clustering approaches are closely related to Gaussian Mixture Models (GMMs) and EM like algorithms have been employed in fuzzy clustering with regularized objective functions. Student's t -mixture models (SMMs) have been proposed recently as an alternative to GMMs, resolving their outlier vulnerability problems. Chatzis et al., [20] 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.

Sato-Ilic et al., [21] 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.

Jiabin Deng et al., [22] proposed an improved fuzzy clustering-text clustering method based on the fuzzy C-Means 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 high-power sample point set, the field radius and weight. Due to the boundary value attribution of the traditional fuzzy C-Means clustering algorithm, the author recommended the edit distance algorithm.

A fuzzy system entirely characterizes one region of the input-output product space S=U×V through a relation expressed by a set of fuzzy rules. Effectively, the fuzzy system establishes a fuzzy map, which assigns for each input fuzzy set in U an output fuzzy set in V. The partition of this product space may be made through the decomposition of the relation. The fuzzy clustering of fuzzy rules, here proposed, as well as clustering of data, leads to a fuzzy partition of the S space. The result is a set of fuzzy sub-systems, one for each cluster that will be conveniently linked in a new structure. Salgado et al., [23] proposed a new recursive clustering algorithm for the partition of a fuzzy system into a hierarchical collaborative structure. The global response of the hierarchical collaborative structure is identical to the input fuzzy system.

Celikyilmaz et al., [24] 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 c-partitioning 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.

With the growing demand for categorical data clustering, a few clustering algorithms with focus on categorical data have recently been developed. However, most of these methods attempt to optimize a single measure of the clustering goodness. Often, such a single measure may not be appropriate for different kinds of datasets. Thus, consideration of multiple, often conflicting, objectives appears to be natural for this problem. Although the authors have previously addressed the problem of multiobjective fuzzy clustering for continuous data, these algorithms cannot be applied for categorical data where the cluster means are not defined. Motivated by this, a multiobjective genetic algorithm-based approach for fuzzy clustering of categorical data is proposed by Mukhopadhyay et al., [25] that encodes the cluster modes and simultaneously optimizes fuzzy compactness and fuzzy separation of the clusters. Moreover, a novel method for obtaining the final clustering solution from the set of resultant Pareto-optimal solutions is proposed. This is based on majority voting among Pareto front solutions followed by k-nn classification.

In order to clustering the temporal or spatial dataset, constraints-equipped version of fuzzy C-Means was proposed in literature. RuiqiongCai et al., [26] 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.

4. Problems and the Directions

One of the fundamental difficulties that arise in several fields, comprising pattern recognition, machine learning and statistics, is clustering. The basic data clustering problem might be defined as finding out groups in data or grouping related objects together. A cluster is a group of objects which are similar to each other within a cluster and are dissimilar to the objects of other clusters. The similarity is typically calculated on the basis of distance between two objects or clusters. Two or more objects present inside a cluster and only if those objects are close to each other based on the distance between them.

The major objective of clustering is to discover collection of comparable objects based on similarity metric. On the other hand, a similarity metric is generally specified by the user according to his requirements for obtaining better results. So far, there is no such technique available which absolutely fits for all applications. Some of the major difficulties concerning the existing available clustering approaches are that they do not concentrate on the entire needs effectively and require huge time complexity in case of clustering a great number of dimensions and bulky data sets.

Efficiency of a particular clustering approach chiefly based on the definition of the distance, means that the measure of distance between the two objects in a particular cluster should be well defined using effective distance measures. Also it is necessary to know about the effect of constraints in clustering the objects. The use of constraints in clustering along with the effective distance measures will definitely provide better clustering results.

There are several approaches available for clustering objects. Out of those K-Means is one of the basic and a standard algorithm that is most commonly used in many applications. This approach clusters objects to a pre-defined number of clusters, which is specified by the user. This approach selects the cluster centroid (center) for each cluster at random. These initial centroids are selected at random in such a way that it should be far away from each other. It is to be noted that these initial centroids completely affect the clustering results, since the entire clustering process is carried out on the basis of these initial centroids. Apart from K-Means, other clustering approaches are, Fuzzy C-Means and Possiblistic C-Means. But these techniques are not suitable for all applications and huge data collections.

Clustering has become a vital area of research in the field of data mining. It has been widely used in various applications. There are various clustering techniques available in the literature. But these conventional systems have lot of limitations. The accuracy of the existing clustering approaches is a major concern. The time utilized for active clustering of data is more if large databases are taken up for clustering. The other problems concerned with the traditional clustering techniques are:

- Traditional Clustering techniques are quite sensitive to initial positions of cluster centers [27]
- Cluster centroids may not be optimal as the conventional algorithms can converge to local optimal solutions [28]
- Mean Squared Error (MSE) value is very high in the conventional techniques [29]
- The conventional clustering do not perform well in vague and large data sets [30]

So, a novel clustering approach is very much necessary to overcome the above given problems.

5. 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. Market basket analysis is one of the main applications of clustering in supermarkets. In marketing, clustering finds groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records. Classification of plants and animals given their features is also a major application area in bioinformatics. In World Wide Web, Document classification and clustering weblog data to discover groups of similar access patterns is an active area of research. Clustering has been one of the most vital techniques in the field of data mining. Recently, clustering is applied in various 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.

References

[1] Fayyad, Usama, Gregory Piatetsky-Shapiro and Padhraic Smyth, “From Data Mining to Knowledge Discovery in Databases”, American Association for Artificial Intelligence- AI Magazine, Pp. 37-54, 1996.
[2] G. Karypis, E.-H Han and V. Kumar, “Chameleon: Hierarchical Clustering using Dynamic Modeling”, IEEE Computer, Pp. 68–75, 1999.
[3] D.H. Fisher. “Conceptual clustering, learning from examples, and inference”, Proc. 4th Int. Workshop on Machine Learning, Irvine, CA, Pp. 38–50, 1987.
[4] Osmar R. Zaïane, “Principles of Knowledge Discovery in Databases Chapter8:DataClustering”,http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html.
[5] A.K. Jain and R.C. Dubes, “Algorithms for Clustering Data” ,Prentice Hall, 1988.
[6] E. Chandra, V.P. Anuradha, “Survey on Clustering Algorithms for Data in Spatial Database Management Systems”, International Journal of Computer Applications vol. 24, no.9, pp.975 – 8887, June 2011.
[7] RuiXu,Donald Wunsch, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005.
[8] A.K. Jain, M.N. Murty, P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 199.
[9] Parimala, Daphne Lopez, N.C. Senthilkumar, “A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases”, International Journal of Advanced Science and Technology, vol. 31, June, 2011.
[10] Don Kulasiri, Sijia Liu, Philip K. Maini and RadekErban, “DifFUZZY: A fuzzy clustering algorithm

for complex data sets" , International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol.1, no.4,pp. 402-417, 2010.

[11] Anil K. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, no.31, pp. 651–666, 2010.

[12] 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.

[13] F. Yang, T. Sun and C. Zhang, "An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization", International Journal on Expert Systems with Applications, Vol. 36, No. 6, Pp. 9847-9852, 2009.

[14] Khan D.M and Mohamudally N, "A multiagent system (MAS) for the generation of initial centroids for k-means clustering data mining algorithm based on actual sample datapoints", 2nd International Conference on Software Engineering and Data Mining (SEDM), Pp. 495–500, 2010.

[15] Chen Zhang and Shixiong Xia, "K-means Clustering Algorithm with Improved Initial Center", Second International Workshop on Knowledge Discovery and Data Mining (WKDD), Pp. 790–792, 2009.

[16] Pardeshi B and Toshniwal D, "Improved k-medoids clustering based on cluster validity index and object density", IEEE 2nd International Advance Computing Conference (IACC), Pp. 379–384, 2010.

[17] Yanfeng Zhang, XiaofeiXu and Yunming Ye, "NSS-AKmeans: 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.

[18] Baraldi A and Blonda P, "A survey of fuzzy clustering algorithms for pattern recognition. II", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 29, No. 6, Pp. 786–801, 1999.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Salgado P and Garrido P.J, "Fuzzy clustering of fuzzy systems", IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, Pp. 2368–2373, 2004.

[24] 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.

[25] Mukhopadhyay A, Maulik U and Bandyopadhyay S, "Multiobjective Genetic Algorithm-Based Fuzzy Clustering of Categorical Attributes", IEEE Transactions on Evolutionary Computation, Vol. 13, No. 5, Pp. 991–1005, 2009.

[26] 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.

[27] P.S. Bradley and U.M. Fayyad, "Refining Initial Points for K-Means Clustering", ACM, Proceedings of the 15th International Conference on Machine Learning, Pp. 91-99, 1998.

[28] Wei-Chuan Liu, Jiun-Long Huang and Ming-Syan Chen, "KACU: k-means with hardware centroid-updating",Emerging Information Technology Conference, 2005.

[29] Shi Yong and Zhang Ge, "Research on an improved algorithm for cluster analysis", International Conference on Consumer Electronics, Communications and Networks (CECNet), Pp. 598–601, 2011.

[30] Weijiang Jiang and Jun Ye, "Decision-making method based on an improved similarity measure between vague sets", IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design (CAID & CD), Pp. 2086–2090, 2009.

Author Profile

L. Divya Sivanandini received Bachelor degree in Chemistry from P.S.G.R Krishnammal College of art and science and MCA degree from GR Govindarajulu School of applied computer technology, Coimbatorein 2007 and 2010 respectively. Currently pursuing Master of Philosophy in Computer Science in Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore under the Guidance of Mr. M. Mohanraj M.C.A., M.Phil., Assistant Professor, Dept Of Computer Applications. Her area of Interest is Data clustering.

Mohanraj M received the BCA degree in Computer Applications from Bharathiar University and MCA degree from Bharathiyar University, Coimbatore in 2003 and 2006 respectively. He received his M. Phil from Periyar University in 2008 and he is doing Doctor of Philosophy in Computer Science and Engineering from Periyar University, Salem. He had seven years of teaching experience. Currently he is working as Assistant Professor in Computer Application Department, Dr.SNS Rajalakshmi College of Arts and Science College, Coimbatore.