Enhancing the Performance of K-Means Clustering by using Fuzzy Partitioning Matrix

Ahtesham Husain Shaikh
P.G. Student
SSBT's College of Engg. & Tech.
Bambhori, Jalgaon M.S. India

Manoj E. Patil
Associate Prof.
SSBT's College of Engg. & Tech.
Bambhori, Jalgaon M.S. India

ABSTRACT
Clustering hast two approaches, Hard clustering and soft clustering. The hard clustering restricts that the data object in the given data belongs to exactly one cluster. The problem with hard K-Means (KM) clustering is that the different initial partitions can result in different final clusters. Soft clustering which also known as fuzzy clustering forms clusters such that data object can belong to more than one cluster based on their membership levels. But sometimes the resulting membership values do not always correspond well to the degrees of belonging of the data. So to overcome the problems in hard Fuzzy K-Means clustering, the improved Fuzzy K-Means (FKM) clustering approach is proposed. The proposed improved Fuzzy K-Means clustering assigns membership to an object inversely related to the relative distance of the object to cluster prototype. Fuzzy K-Means clustering assigns membership levels which indicate the degree to which the data elements belong to the clusters, and then using them to assign data object to one or more clusters. These indicate the strength of the association between that data object and a particular cluster. The proposed work also compares the execution time and required memory of Proposed Fuzzy K-Means (FKM) to that of existing Fuzzy K-Means clustering.

Keywords
Fuzzy clustering, Fuzzy Partition Matrix

1. INTRODUCTION
Clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data. It groups the data objects according to measured or perceived intrinsic characteristics or similarity. Cluster analysis does not use category labels that tag objects with prior identifiers, i.e., class labels.

The absence of category information distinguishes data clustering (unsupervised learning) from classification or discriminate analysis (supervised learning).

The clustering in data mining becomes very difficult because of very large datasets with many attributes of different types. This causes to have unique computational requirements on appropriate clustering algorithms. The main concern for most of clustering algorithms is their need to know the number of clusters for which to look.

Since the clustering is an unsupervised way of grouping, the user has no previous knowledge about the actual number of clusters. Apparently, dividing the dataset into smaller or larger clusters will result in merging some separate clusters or breaking down some compact ones. The process of finding an optimal number of clusters is called cluster validity.

In order to achieve the main aim of fuzzy k-means clustering, the drawbacks of traditional k-means clustering are studied. k-means clustering clusters the data in a crisp sense which results into empty clusters.

Whereas, the proposed Fuzzy K-Means (FKM) clustering uses the membership partition matrix grades in order to express ambiguity in the assignment of data point to clusters. The proposed partition based fuzzy k-means employs fuzzy measures as the basis for membership matrix calculation and for cluster centers identification. The fuzzy measures applied to clustering helps to improve the results of fuzzy k-means clustering.

2. RELATED WORK
Fuzzy logic introduced by Zadeh [12] may be viewed as an attempt at formalization of two remarkable human capabilities. The capability to perform a wide variety of physical and mental tasks without any measurements and any computations. In many real world application areas, knowledge is represented by in terms of imprecise linguistic words from a natural language. A linguistic variable means a variable whose values are words or sentences in a natural language or artificial language. For example, honesty is linguistic variable. The linguistic values of this variable can be extremely honest, not honest, sometimes honest, and very honest.

Fuzzy logic is the way of representing and manipulating data that is not exact, but rather uncertain [11]. Uncertainty can be manifested in many forms: it can be fuzzy (not sharp, unclear, imprecise, approximate), it can be vague (not specific, amorphous), it can be ambiguous (too many choices, contradictory), it can be of the form of ignorance (dissonant, not knowing something), or it can be a form due to natural variability (conicting, random, chaotic, unpredictable).

3. LITERATURE SURVEY
Figure 1 shows the structure of literature survey. There are two main domains of data mining Clustering and classification. The clustering is unsupervised learning and classification is supervised learning. Clustering is divided into two categories namely hard clustering and fuzzy clustering. Hard clustering clusters the data in a crisp sense. It means each data object can be a member of one and only one cluster at a time. In Hard clustering there is always at least one object in each cluster. However, the empty clusters can be obtained if not a single object is allocated to a cluster during the assignment. The Fuzzy clustering assigns the data object to more than one cluster at a time.

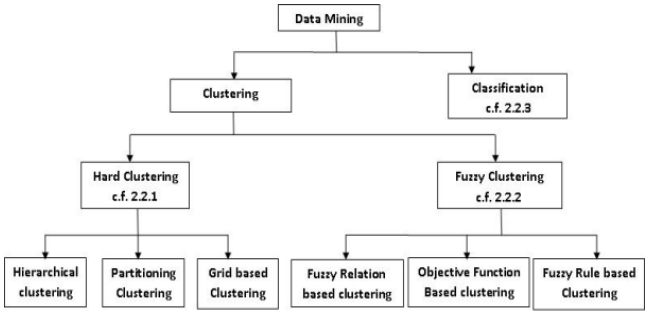

Figure 1: Literature Survey

3.1 Hard Clustering

Hard clustering is also known as crisp clustering. Crisp clustering allocate each data pattern (data object) of given input to a single cluster. Thus in hard clustering, each data pattern (data object) belongs to only one cluster. Farley and Raftery, in [17], suggested dividing the clustering methods into two main groups: partitioning and hierarchical methods. Han and Kamber, in [1], suggested categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods.

3.2 Partitioning Clustering

Partitioning clustering [18] directly divides data objects into some pre-specified number of cluster. The checking for all possible cluster is computationally impractical, certain greedy heuristics are used in the form of iterative optimization of cluster. Researchers have suggested several partitioning clustering approaches viz., K-Means Clustering, K-Medoid Clustering, Relocation Algorithm and Probabilistic Clustering etc.

K. Tapas et al., in [19], have proposed K-means clustering. K-Means clustering is a method commonly used to automatically partition a data set into clusters (K). Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters. Each cluster is characterized by its centre point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. The K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co-ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters. The input parameters of the clustering algorithm are the number of clusters that are to be found along with the initial starting point values. When the initial starting values are given, the distance from each sample data point to each initial starting value is found. Then each data point is placed in the cluster associated with the nearest starting point. After all the data points are assigned to a cluster, the new cluster centroids are calculated. For each factor in each cluster, the new centroid value is then calculated. The new centroids are then considered as the new initial starting values. This process continues until no more data point changes or until the centroids no longer move. In K-Means data object can belong precisely to only one cluster during clustering process. This can be too restrictive while clustering high dimensional data expressed in multiple conditions. Han and Kamber, in [1], have proposed K-medoid Clustering. In the K-medoid clustering a cluster is represented by one of its points called medoid. A medoid is the centrally located data point. When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid. Every time a new medoid is selected, the distance between each object and its newly selected cluster center has to be recomputed. Because there could be obstacles between two objects, the distance between two objects may have to be derived by geometric computations. The computational cost can get very high if a large number of objects and obstacles are involved. Representation by k-medoids has two advantages [9]. First, it presents no limitations on attributes types, and, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and therefore, it is lesser sensitive to the presence of outliers. P. Berkhin, in [3], have proposed Relocation Algorithms. The relocatopon algorithms iteratively reallocate points between the k clusters. The points are reassigned based on the local search algorithm. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The three changeable elements of the general relocation algorithm are initialization, reassignments of the data points into clusters and update of the cluster parameters. These algorithms builds the high quality clusters due to iterative approach. The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as iterative relocation. I. V. Cadez et al., in [20], have proposed Probabilistic Clustering. In the probabilistic approach, data is considered to be a sample independently drawn from a mixture model of several probability distributions. SO the clusters are associated with the corresponding distributions parameters such as mean and variance.

3.3 Hierarchical Clustering

IndiraPriya and Ghosh, in [21], have proposed hierarchical clustering. Hierarchical clustering creates a hierarchical decomposition of the given set of data objects. It builds a cluster hierarchy, a tree of cluster, also known as a dendrogram. It represents a sequence of nested cluster which constructed top-down or bottom-up. The root of the tree represents one cluster, containing all data points, while at the leaves of the tree, there are n clusters, each containing one data point. By cutting the tree at a desired level, a clustering of the data points into disjoint groups is obtained. A hierarchical clustering is used to find data on different levelsof dissimilarity. A hierarchical clustering can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

3.4 Density-Based Algorithms

P. Berkhin, in [3], have proposed density-based algorithms. Density-based algorithms are capable of discovering clusters of arbitrary shapes. These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter.

3.5 Model-Based Clustering

Han and Kamber, in [1], have proposed model-based clustering methods. Model-based clustering hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on

standard statistics, taking noise or outliers into account and thus yielding robust clustering methods. Expectation-Maximization (EM) is an algorithm that performs expectation-maximization analysis based on statistical modeling. Cobweb is a conceptual learning algorithm that performs probability analysis and takes concepts as a model for clusters. Self-Organizing feature Map (SOM) is a neural network-based algorithm that clusters by mapping high dimensional data into a 2-D or 3-D feature map, which is also useful for data visualization.

3.6 Grid-Based Clustering

Grid-based clustering [1] quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. Some typical examples of the grid-based approach include STING (Statistical Information Grid) , which explores statistical information stored in the grid cells; WaveCluster, which clusters objects using a wavelet transform method; and CLIQUE (CLustering InQUEst) , which represents a grid-and density-based approach for clustering in high-dimensional data space. Grid based clustering create a grid structure by partitioning the data space into a finite number of non-overlapping cells then calculate the cell density for each cell. After calculating density grid based clustering sort the cells according to their densities. Cluster centers are identified and all neighbour cells are traversed.

N. H. Park and W. S. Lee, in [23], presented statistical grid-based clustering over data streams. A data stream is a large unbounded sequence of data elements continuously generated at a rapid rate. The approach used is statistical grid-based approach for clustering data elements of data streams. First, the multidimensional data space of a data stream is partitioned into a set of mutually exclusive equal size initial cells. When the support of a cell becomes high enough, the cell is dynamically divided into two mutually exclusive intermediate cells based on its distribution statistics. A cluster of a data streams is group of adjacent dense unit cells.

3.7 Fuzzy Clustering

Fuzzy clustering is the synthesis between the fuzzy logic and clustering which is the requirement of modern computing [12]. The aim of fuzzy clustering is to model the ambiguity within the unlabeled data objects efficiently. Every data object is assigned a membership to represent the degree of belonging to certain class. The requirement that each object is assigned to only one cluster is relaxed to weaker requirement in which the object can belong to all of the clusters with a certain degree of membership. Thus it assigns degrees of membership in several clusters to each input pattern. A fuzzy clustering can be converted to a hard clustering by assigning each pattern to cluster with the largest measure of membership. Soft clustering is categorized in three categories [15]: Fuzzy relation based clustering, fuzzy rule based clustering, objective function based clustering.

3.8 Fuzzy Relation Based Clustering

M. S. Yang, in [15], have proposed fuzzy relation based clustering. Fuzzy relation based clustering includes an N-step procedure by using the composition of fuzzy relations beginning with a reflexive and symmetric fuzzy relation R in X. The data set is partitioned into the number of cluster by equivalence relation. G. S. Liang et. al., in [24], have introduced cluster analysis based on fuzzy equivalence relation. The approach used is the distance measure between two trapezoidal fuzzy numbers is used to aggregate subjects linguistic assessments. The distance measure is used to characterize the interobjects similarity. The linguistic assessment is for attributes ratings to obtain the compatibility relation. Then a fuzzy equivalence relation based on fuzzy compatibility relation is constructed.

4. PROPOSED APPROACH

The proposed solution focuses on text clustering using fuzzy logic based clustering in order to facilitate and improve effectiveness in an conventional hard clustering approach.

Fuzzy clustering is a partition based clustering scheme and is particularly useful when there are no apparent clear groupings in the data set [34]. Partitioning schemes provide automatic detection of cluster boundaries and in case of fuzzy clustering, these cluster boundaries overlap. Every individual data entity belongs to not one but all the clusters with varying degrees of membership.

The proposed system preprocess the data, then the preprocessed data is given as an input to the conventional hard clustering algorithm and proposed fuzzy portioning algorithm. Finally, the cluster formation is done and the results of both the hard clustering and fuzzy clustering are compared.

Hard partition is insufficient to represent many real situations. Therefore, a fuzzy clustering method is offered to construct clusters with uncertain boundaries. Hence, this method allows that one object belongs to some overlapping clusters to some degree.

Fuzzy clustering is a partition based clustering scheme and is particularly useful when there are no apparent clear groupings in the data set [34]. Partitioning schemes provide automatic detection of cluster boundaries and in case of fuzzy clustering, these cluster boundaries overlap. Every individual data entity belongs to not one but all the clusters with varying degrees of membership.

4.1 Architecture

The architecture of the proposed system is shown in Figure 2. Input to the proposed system is the text data set [36]. The data set contains text files. The data preprocessing consists of the stemming, removal of stop words, feature selection, create vector of each object as shown in Figure 3.2. After data preprocessing each text file is given as an input to the hard clustering and fuzzy clustering.

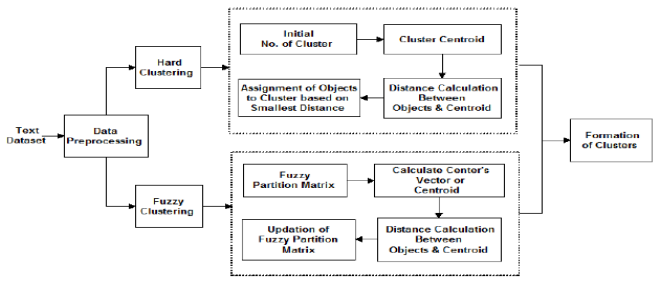

Figure 2: Architecture of Proposed System

In fuzzy clustering, a data object will have an associated degree of membership for each cluster, indicating the strength of its association in that cluster. It iteratively update the membership values of a data object with the pre-defined

number of clusters. Thus, a data object can be the member of all clusters with the corresponding membership values. The process of calculation of cluster centers and the assignment of points to these centers is repeated until the cluster centers stabilize [37].

4.2 Proposed Fuzzy Partitioning Matrix

Classically, the clustering has been based on the disjointness condition that no two data objects belong to same cluster. Hard clustering algorithms partitions the data set X into specified number of mutually exclusive subsets of X. However, in real data sets a data object may belong to various clusters. Hence, such situations require weakening of disjointness condition. In fuzzy clustering, an object can belong to several clusters simultaneously, with different degrees of membership. However, Fuzzy K-Means still uses a cost function that is to be minimized while trying to partition the data set.

4.3 Preprocessing of Input Email Data set

Stop Words Removal

Algorithm 1 presents removal of stop words. A stop word is defined as a term which is not thought to convey any meaning as a dimension in the vector space (i.e. without context). The standard set of stop words provide a valid set of words to prune.

Feature Selection based on Document Frequency

Algorithm 2 shows the document frequency based feature selection. It first counts the TermFrequency (TF). In TF the total no. of occurrence of term is counted. Then the Inverse Document Frquency (IDF) is counted.

4.4 Proposed Fuzzy K-Means Clustering

The Fuzzy-k-Means Procedure. The clusters produced by the k-means procedure are sometimes called "hard" or "crisp" clusters, since any feature vector x either is or is not a member of a particular cluster.

The Algorithm is of proposed Fuzzy K-Means. The major process of Proposed Fuzzy K-Means (FKM) is mapping a given set of representative vectors into an improved one through partitioning data points. It begins with a set of initial cluster centers and repeats this mapping process until a stopping criterion is satisfied. It is supposed that no two clusters have the same cluster representative. In the case that two cluster centers coincide, a cluster center should be perturbed to avoid coincidence in the iterative process.

5. RESULTS AND DISCUSSION

The Enron Email Data set [36] is used for experimenting the proposed work. The data set contains numerous E-mail messages. The E-mail consists of three types of features: unstructured text, categorical text, and numeric data. Unstructured text in email consists of fields like the subject and body, which allow for natural language text of any kind. Categorical text includes fields such as "to" and "from". These differ from unstructured text fields in that the type of data which can be used in them is very well defined. Numeric Data in email includes such features as the message size, number of recipients.

Table shows data structure for stemming. Preprocessing contains stemming and stopword removal. Stemming attempts to remove the differences between in ected forms of a word, in order to reduce each word to its root form. Stopwords are words which have very little informational content. These are words such as:and, the, of, it, as, may, that, a, an, of, off, etc. One way to do stemming is to store a table of all index terms and their stems

5.1 Experimental Setup

The proposed system is evaluated using JAVA NetBeans IDE 8.0.2 on Windows XP operating system. NetBeans is more than Just an editor. For experimenting the proposed work Enron Email Data set [36] is used. The data set contains numerous E-mail messages.

Table shows the input parameter for proposed system. The table shows Number of Samples, All Unique Terms, TFIDF (Term Frequency-Inverse Document Frequency) Weight, Number of Iterations (NOI), Fuzziness Factor, and the last column is of Number of Clusters (K).

Table 1: Input Parameters

No of Samp.	U.T.	R.T.	TFIDF	NOI	Fuzzy Factor	K
16	1338	275	16	3	0	3
79	2817	883	79	5	1	6
145	2163	766	145	7	2	9
240	4971	1622	240	10	3	12
350	5783	1940	350	15	4	15
971	5442	2121	971	20	5	18

5.2 Experimental Results

The experimental results show the result analysis of proposed work. Table 4.7 shows the result analysis of proposed Fuzzy K-Means based on the parameters such as Number of Samples, K-Value (no. of clusters), Number of Iterations (NOI), Fuzziness Factor and Maximum number of similar objects between clusters formed by Proposed Fuzzy K-Means (FKM) and hard K-Means (KM). The Fuzziness Factor affects the membership distribution of an object. It simply used to control how much clusters are allowed to overlap. The higher the value of Fuzziness Factor, the larger the overlap between clusters. In other words, the higher the fuzziness factor the algorithm uses, a larger number of data objects will fall inside a fuzzy band.

Table 2: Result Analysis of Proposed Fuzzy K-Means

No of Samples	K-Value	NOI	Fuzziness Factor	Max no of Similar Obj. bet. FKM and HKM
79	4	5	1	27
100	5	7	2	31
150	6	10	4	37
250	5	14	7	43
350	7	15	12	47

In Table, when No.of Samples are 79, K=4, NOI=5 and fuzziness Factor=1, the max imum number of similar objects between Proposed FKM and KM are 27. The maximum number of similar objects between Proposed FKM and KM is depends on the Fuzziness Factor. When No. of Samples are 100, K=5, NOI=7 and fuzziness Factor=2, the maximum number of similar objects between Proposed FKM and KM are 31. It shows that as the Fuzziness Factor increases the

maximum number of similar objects between proposed FKM and KM.

Table 3: Similarity of Objects between Proposed FKM and Hard KM when Fuzziness Factor is 1

	KM Cluster 1	KM Cluster 2	KM Cluster 3	KM Cluster 4
FKM Cluster 1	1	2	8	1
FKM Cluster 2	4	1	5	27
FKM Cluster 3	6	4	0	13
FKM Cluster 4	0	3	12	5

Consider the KM Cluster 1 and FKM Cluster 1 in Table 4.8, it shows that there is 1 similar objects. KM Cluster 1 is compared with FKM Cluster 2, it shows there are 4 similar objects.

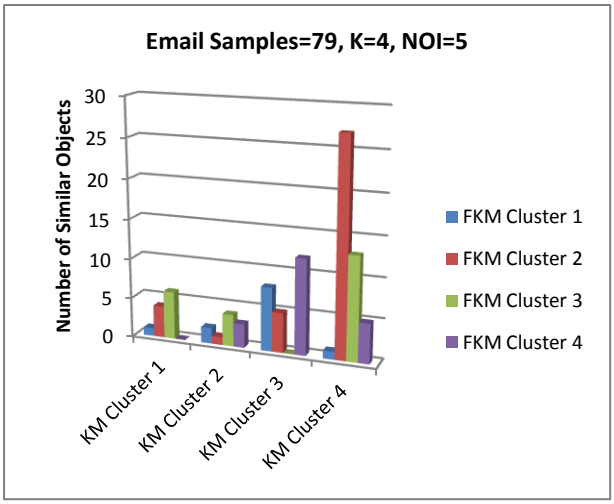

Figure: 3 Graph for similarity per cluster when Fuzziness Factor is 1.

K-Means (KM) which is based on Fuzziness Factor. The X axis represents the Fuzziness Factor and Y axis represents Number of Similar objects. It shows that when Fuzziness Factor is 0, the maximum 20 similar objects are clustered into 7 clusters of both Proposed FKM and KM. When Fuzziness Factor is 1, the maximum 29 similar objects are clustered into 7 clusters of both Proposed FKM and KM.

Table 4: Object Similarity based on Fuzziness Factor

Samples	NOI	K	Fuzzy Factor	Sim Obj in FKM and KM
30	15	7	0	20
			2	41
			4	47
			6	59
			8	70
			10	81
			15	111

Table shows the similarity of objects between Proposed Fuzzy K-Means (FKM) and hard K-Means (KM) when Fuzziness Factor is 1. 1st cluster of KM is compared with the1st, 2nd, 3rd and 4th cluster of Proposed FKM respectively.

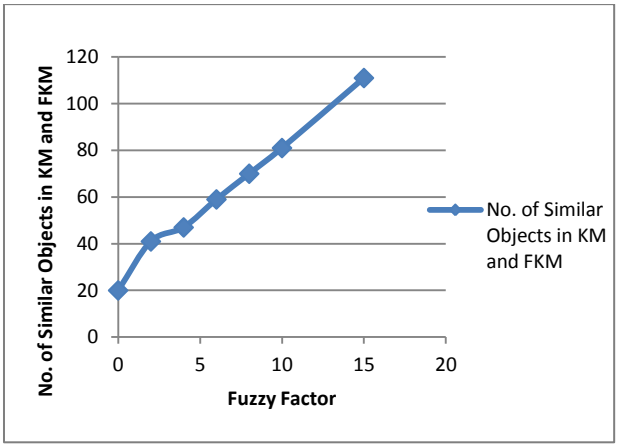

Figure 4: Object Similarity based on Fuzziness Factor

6. CONCLUSION AND FUTURE WORK

The Proposed FKM formulate the objective function in terms of improving the membership assignments of an object. It assigns fuzzy memberships to data object and updates the centre of cluster according to the assigned memberships. The assigned memberships play a role as weight values which represent the degree to which data object belongs to more than one clusters. The degree of belongingness depends on the selection of Fuzziness Factor. The Proposed Fuzzy K-Means significantly differ depending on the choice of Fuzziness Factor. In hard K-Means a set of k initial cluster centers is chosen arbitrarily and each object is then assigned to the center closest to it, and the centers are recomputed. This is repeated until the process stabilizes which takes more execution time and memory. On the other hand, in Proposed Fuzzy K-Means approach though it assigns membership to an object which is inversely related to the relative distance of the object to cluster centre, the Proposed Fuzzy K-Means approach takes less execution time and requires less memory than that of hard K-Means.

7. REFERENCES

[1] J. Han and M. Kamber, “Data mining: concepts and techniques," 2001.

[2] C. C. Aggarwal and C. K. Reddy, “Data clustering: algorithms and applications”.

[3] P. Berkhin, “Survey of clustering data mining techniques," San Jose, CA, 2002,

[4] O. M. Jafar and R. Sivakumar, “A comparative study of hard and fuzzy data clustering algorithms with cluster validity indices," in Proceedings of the Elsevier International

[5] J. Daxin, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: a survey", IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370.

[6] A. Baraldi and P. Blonda, “A survey of fuzzy clustering algorithms for pattern recognition," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 29, no. 6, pp. 778-785, october 1998.

[7] J. Hu, C. Xiong, J. Shu, X. Zhou, and J. Zhu, "A novel text clustering method based on tgsom and fuzzy k-means" , in Proceedings of the 2009 First International Workshop on Education Technology and Computer Science - Volume 01, ser. ETCS '09, 2009, pp. 26-30.

[8] C. Wu, C. Ouyang, L. nan Chen, and L. Lu, "A new fuzzy clustering validity index with a median factor for centroid based clustering", IEEE Transactions on Fuzzy Systems, vol. 23, no. 3, June 2015.

[9] L. Rokach and O. Maimon, "Clustering methods," in Data mining and knowledge discovery handbook. Springer, 2005, pp. 321-352.

[10] N. A. M. Isa, S. Salamah, U. K. Ngah et al., "Adaptive fuzzy moving k-means clustering algorithm for image segmentation," IEEE Transactions on Consumer Electronics, vol. 55, no. 4, pp. 2145-2153, 2009.

[11] T. J. Ross, Fuzzy logic with engineering applications, 2nd ed. John Wiley & Sons, 2009.

[12] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, pp. 338-353, 1965.

[13] Zadeh, "Is there a need for fuzzy logic?" Information sciences, vol. 178, no. 13, pp. 2751-2779, 2008.

[14] L. Zadeh, C. Negoita, and H. Zimmermann, "Fuzzy sets as a basis for a theory of possibility," Fuzzy sets and systems, vol. 1, pp. 3-28, 1978.

[15] M. S. Yang, "A survey of fuzzy clustering," Mathematical and Computer modelling, vol. 18, no. 11, pp. 1-16, 1993.

[16] Q. Ni, Q. Pan, H. Du, C. Cao, and Y. Zhai, "A novel cluster head selection algorithm based on fuzzy clustering and particle swarm optimization," IEEE/ACM Transactions on Computational Biology and Bioinformatics, no. 99, pp. 1{9, 2015.

[17] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? Answers via model-based cluster analysis," The computer journal, vol. 41, no. 8, pp. 578-588, 1998.

[18] S. Ayramo and T. Karkkainen, "Introduction to partitioning based clustering methods with a robust example," Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, 2006.

[19] K. Tapas, D.M.Mount, N. Netanyahu, C. Piatko, R. Silverman, and A.Y.Wu, "An efficient k-means clustering algorithm: analysis and implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, July 2002.

[20] I. V. Cadez, S. Ga_ney, and P. Smyth, "A general probabilistic framework for clustering individuals and objects," in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000, pp. 140-149.

[21] P. IndiraPriya and D. Ghosh, "A survey on different clustering algorithms in data mining technique," International Journal of Modern Engineering Research (IJMER), vol. 3, no. 1, pp. 267-274, 2013.

[22] M. W. Berry and M. Castellanos, Survey of Text Mining:Clustering, Classification and Retrieval, 2nd ed. Springer, 2007.

[23] N. H. Park and W. S. Lee, "Statistical grid-based clustering over data streams," ACM SIGMOD Record, vol. 33, no. 1, pp. 32-37, 2004.

[24] G.-S. Liang, T.-Y. Chou, and T.-C. Han, "Cluster analysis based on fuzzy equivalence relation," European Journal of Operational Research, vol. 166, no. 1, pp. 160-171, June 2004.

[25] M.-S. Yang and H.-M. Shih, "Cluster analysis based on fuzzy relations," Fuzzy Sets and Systems, vol. 120, no. 2, pp. 197{212, 2001.

[26] E. G. Mansoori, "Frbc: a fuzzy rule-based clustering algorithm," IEEE Transactions on Fuzzy Systems, vol. 19, no. 5, pp. 960-971, 2011.

[27] M. Delgado, A. F. G_omez-Skarmeta, and F. Martin, \A fuzzy clustering-based rapid prototyping for fuzzy rule-based modeling," IEEE Transactions on Fuzzy Systems, vol. 5, no. 2, pp. 223{233, 1997.

[28] Y. Lu, T. Ma, C. Yin, X. Xie, W. Tian, and S. Zhong, \Implementation of the fuzzy c-means clustering algorithm in meteorological data," International Journal of Database Theory and Application, vol. 6, no. 6, pp. 1-18, 2013.

[29] P. Lingras and G. Peters, "Applying rough set concepts to clustering," in Rough Sets: Selected Methods and Applications in Management and Engineering. Springer, 2012,

[30] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," IEEE Transactions on Fuzzy Systems, vol. 3, no. 3, pp. 370-379, 1995.

[31] P. Maji and S. Paul, "Rough-fuzzy clustering for grouping functionally similar genes from microarray data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 2, pp. 286-299, March 2013.

[32] O. Sutton, "Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction," University lectures, University of Leicester, 2012

[33] R.-P. Li, M. Mukaidono, and I. B. Turksen, "A fuzzy neural network for pattern classification and feature selection," Fuzzy Sets and Systems, vol. 130, no. 1, pp. 101-108,

[34] L. Zhu, F.-L. Chung, and S. Wang, "Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 39, no. 3, pp. 578-591, 2009.

[35] Y. Karali, D. Kodamasingh, and R. L. H. Behera, "Hard and fuzzy clustering algorithms using normal distribution of data points: a comparative performance analysis," in International Journal of Engineering Research and Technology, vol. 2, no. 10 (October-2013).

[36] "Enron email dataset," https://www.cs.cmu.edu/ ./enron/ [Accessed on: May 2015].

[37] W. Pedrycz and H. Izakian, "Cluster-centric fuzzy modeling," IEEE Transactions on Fuzzy Systems, vol. 22, no. 6, pp. 1585-1597, December 2014.

[38] M. F. Porter, \An algorithm for suffix stripping," Program, vol. 14, no. 3, pp. 130{137, 1980.

[39] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 3, 2003, pp. 1661-1666.

[40] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," journal of computing, vol. 1, no. 1, pp. 7-11, December 2009.

[41] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," International Journal of Engineering Research & Technology (IJERT), vol. 1, no. 6, pp. 1-6, 2012.

[42] A. K. Murugesan and B. J. Zhang, "A new term weighting scheme for document clustering," in 7th International Conference on Data.

[43] J. W. Reed, Y. Jiao, T. E. Potok, B. Klump, M. T. Elmore, A. R. Hurson et al., "Tf icf: A new term weighting scheme for clustering dynamic data streams," in proceedings of 5th IEEE International Conference on Machine Learning and Applications, 2006.

[44] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.