Probability of Fuzzy Clustering for betterment the Purpose Function of K-Means Cluster

Mrs. Pooja R. Kotwal
Computer Engineering
Sandip Institute of Technology and Research Centre Nashik, Maharashtra, India
Mr.Sunil M. Kale
Computer Engineering
Sandip Institute of Technology and Research Centre Nashik, Maharashtra, India
Mrs.Bharati A.Patil
Computer Engineering
Sandip Institute of Technology and Research Centre Nashik, Maharashtra, India
Mr.Pradnyesh J.Bhisikar
Computer Engineering
Sandip Institute of Technology and Research Centre Nashik, Maharashtra, India

ABSTRACT

Clustering is a separation of data particulars into groups of similar objects. Each group, called cluster, consists of objects that are analogous between themselves and different to objects of other groups. The k- Means clustering system is one of the classical and simplest styles for data clustering. It's one of the most widely used styles in practical executions because of its simplicity. But intermittently the performing class values don't always correspond well to the degrees of belonging of the data, so to overcome the problems in hard K-means clustering, the Fuzzy K- Means clustering approach is proposed. fuzzy clustering forms clusters analogous that data object can belong to further than one cluster grounded on their class situations, In the being system deduction structure ideal function is used, it introduced saw tooth nature in objective function. In this paper feasibility of fuzzy partition matrix of objective function in k-means clustering is proposed, it provides smoothness in aphorism tooth nature in objective function, which is main reason for the perfecting the objective function.

Keywords: fuzzy clustering, fuzzy partition matrix

1. Background, Motivation and Objective

Data/Information clustering is one of the important data mining system same class and supreme diversity between different classes. The large number of data objects greatly increases the challenges of comprehending and interpreting the performing accumulation of data. A first step toward addressing the challenge is the use of clustering. Clustering is a common expressive task in which one seeks to identify a finite set of orders or clusters to describe the data. It groups the data objects according to measured or perceived natural characteristics or similarity. Cluster analysis doesn't use order markers that tag objects with previous identifiers. The absence of order information distinguishes data clustering from bracket or distinguishes analysis.

The clustering in data retrieval becomes veritably delicate because of veritably large datasets with numerous attributes of different types.

This causes to have unique computational conditions on applicable clustering algorithms. The main concern for utmost 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. Supposedly, dividing the dataset into lower or larger clusters will affect in incorporating some separate clusters or breaking down some compact bones

.The process of chancing an optimal number of clusters is called cluster validity. In order to achieve the main end of fuzzy k- means clustering, the downsides of traditional k- means clustering are called k- means clustering .Actually clusters the data in a manner which results into empty clusters.

Whereas, the proposed Fuzzy K- Means clustering uses the class partition matrix grades in order to express darkness in the assignment of data point to clusters. The proposed partition grounded fuzzy k-means describes fuzzy measures as the base for class matrix computation and for cluster centres identification. The fuzzy measures applied to clustering helps to ameliorate the results of fuzzy k-means clustering.

2. 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.

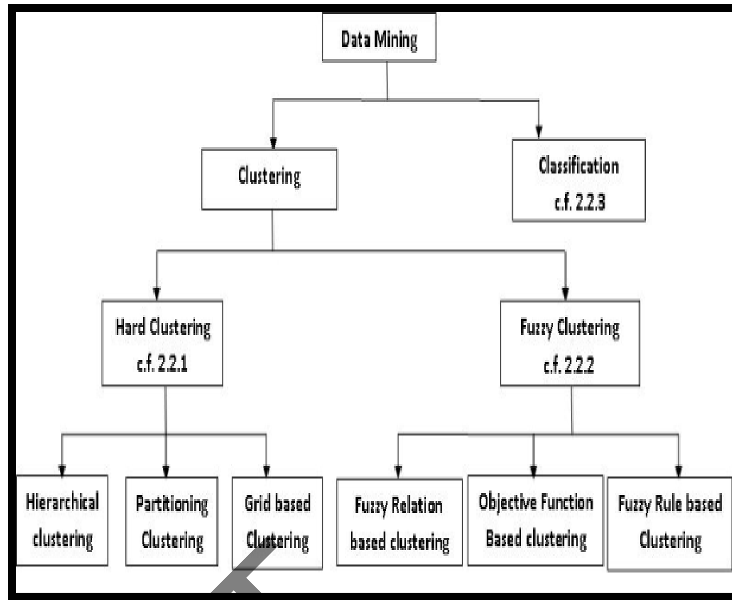

Figure 1: Literature Survey

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.

Fuzzy logic introduced by Zadeh [12] may be viewed as an attempt at formulation 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 the information that is not exact, but rather uncertain [11]. Uncertainty can be manifested in many forms: it can be fuzzy, it can be indistinguishable , 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.

3.2 Partitioning Clustering

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

Tapas et al., in [19], have proposed K- means clustering. K- Means clustering is a system generally 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 utmost of the times don't actually represent the spatial distances. In general, the only result to the problem of chancing global minimum is total choice of starting points. The K- Means clustering algorithm finds the asked number of distinct clusters and their centroids. A centroid is the point whose co-ordinates are attained by means of calculating the normal 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 set up along with the original starting point values. When the original starting values are given, the distance from each sample data point to each original starting value is set up. also 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 also calculated. The new centroids are also considered as the new original starting values. This process continues until no further 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 named, clusters are defined as subsets of points near to separate medoids, and the objective function is defined as the averaged distance or another diversity measure between a point and its medoid. Every time a new medoid is named, the distance between each object and its recently named cluster center has to be recomputed. Because there could be obstacles between two objects, the distance between two objects may have to be deduced by geometric calculations. The computational cost can get veritably 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 mandated by the position of a predominant bit of points inside a cluster and thus, it's lower sensitive to the presence of outliers.P. Berkhin, in [3], have proposed Relocation Algorithms. The algorithms iteratively reallocate points between the k clusters. The points are reassigned grounded on the original hunt algorithm. It also uses an iterative relocation fashion that attempts to ameliorate the partitioning by moving objects from one group to another. The three changeable rudiments of the general relocation algorithm are initialization, reassignments of the data points into clusters and update of the cluster parameters. These algorithms build the high quality clusters due to iterative approach. The process of iteratively reassigning objects to clusters to ameliorate the partitioning is referred to as iterative relocation.[20].So the clusters are associated with the corresponding distributions parameters such as mean and variance.

3.3 Hierarchical Clustering

Indira Priya and Ghosh, in (21), have proposed hierarchical clustering. Hierarchical clustering creates a hierarchical corruption of the given set of data objects. It builds a cluster scale, a tree of cluster, also known as a dendrogram. It represents a sequence of nested cluster which constructed top-down or nethermost- 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 asked position, a clustering of the data points into disjoint groups is attained. A hierarchical clustering is used to find data on different levels of diversity. A hierarchical clustering can be classified as being either agglomerative or divisive, grounded on how the hierarchical corruption 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 neighbourhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighbourhood exceeds some parameter.

3.5 Model-Based Clustering

Han and Kamber, in [1], have proposed model-based clustering methods. Model-based clustering hypothesizes 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 modelling.

3.6 Grid-Based Clustering

One of the most important ways during a perpetration is assessing the quality of your data. When the data has been described and validated, the data is migrated to the new system. This step involves setting up new databases, mapping database fields between systems, and transferring the data.

Grid- grounded 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. The main advantage of this approach is its fast processing time, which is generally independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. Some typical exemplifications of the grid- grounded approach include STING, which explores statistical information stored in the grid cells; Wave Cluster, which clusters objects using a sea transfigure system; and crowd which represents a grid- and viscosity- grounded approach for clustering in high-dimensional data space. Grid grounded clustering produce a grid structure by partitioning the data space into a finite number of non-overlapping cells also calculate the cell viscosity for each cell. After calculating viscosity grid grounded clustering sort the cells according to their consistence. Cluster centres are linked and all neighbour cells are covered.

H. Park and W.S. Lee, in [23], presented statistical grid- grounded clustering over data aqueducts. A data sluice is a large unbounded sequence of data rudiments continuously generated at a rapid-fire rate. The approach used is statistical grid- grounded approach for clustering data rudiments of data aqueducts. First, the multidimensional data space of a data sluice is partitioned into a set of mutually exclusive equal size original cells.

When the support of a cell becomes high enough, the cell is stoutly divided into two mutually exclusive intermediate cells grounded on its distribution statistics.

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, and 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 subject's linguistic assessments. The distance measure is used to characterize the inter objects 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.

3. FEASIBILITY OF FUZZY CLUSTERING FOR IMPROVING THE OBJECTIVE FUNCTION OF K-MEANS CLUSTERING

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 pre-process the data, then the pre-processed 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.

Conditions for a fuzzy context partition matrix, are suggested by relaxing the constraint, this constraints suggest that each target (outlier) word is assigned to at least one of the fuzzy sense cluster with membership greater than zero

$$\mu_{ik} \in [0,1],\ 1 \le i \le c,\ 1 \le k \le N,$$

$$\exists i, \mu_{ik} > 0, \forall k$$

$$0 < \sum_{k=1}^{N} \mu_{ik} < N, 1 \le i \le c$$

FCM calculates the distance between cluster center to the data point and assigns flexible membership to each data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Summation of membership of each data point should be equal to one. FCM uses fuzzy portioning approach, fuzzy partitioning is carried out through the iterative procedure that updates the membership μ_{ij} and cluster center c_j by equations 5.1 d and 5.1 e respectively.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

$$, \quad \forall_j = 1,2, \dots, c$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$$

Where
n = number of data point
C_j = is the j^{th} cluster center
m = fuzziness index m $\in$ [1, ∞]
c = number of cluster center
μ_{ij} = member of i^{th} data to j^{th} cluster
|| || = Euclidean distance

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].

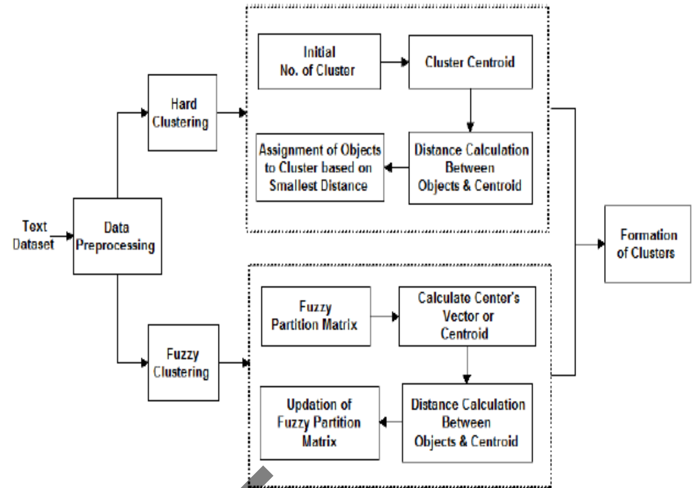

Figure 2: Architecture of Proposed System

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 centres and the assignment of points to these centres is repeated until the cluster centres stabilize [37].

Discussions and Conclusions

The Proposed Fuzzy K-Means formulates the objective function in terms of perfecting the class assignments of an object. It creates the cluster of similar chunks or packets. It assigns fuzzy enrollments to data object and updates the centre of cluster according to the assigned enrollments. The assigned enrollments play a part as weight values which represent the degree to which data expostulate belongs to further 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 original cluster centres is chosen arbitrarily and each object is also assigned to the center closest to it, and the centres are recomputed. This is repeated until the process stabilizes which takes further prosecution time and memory. On the other hand, in Proposed Fuzzy K-Means approach though it assigns class to an object which is equally related to the relative distance of the object to cluster centre.

4. BIBLIOGRAPHY

[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.