An Analysis of Fuzzy Clustering Methods

Virender Kumar Malhotra, Harleen Kaur, M.Afshar Alam
Department of Computer Science
Hamdard University
New Delhi, India

ABSTRACT
Fuzzy logic is an organized and mathematical method of handling inherently imprecise concepts through the use of membership functions, which allows membership with a certain degree. It has found application in numerous problem domains. It has been used in the interval [0, 1] fuzzy clustering, in pattern recognition and in other domains. In this paper, we introduce fuzzy logic, fuzzy clustering and an application and benefits. A case analysis has been done for various clustering algorithms in Fuzzy Clustering. It has been proved that some of the defined and available algorithms have difficulties at the borders in handling the challenges posed in collection of natural data. An analysis of two fuzzy clustering algorithms namely fuzzy c-means and Gustafson-Kessel fuzzy clustering algorithm has been analyzed.

Keywords
Fuzzy logic, Fuzzy clustering Algorithms, Fuzzy C-Means (FCM), Gustafson-Kessel (GK)

1. INTRODUCTION
Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth-truth values between 'completely true' and 'completely false'. Fuzzy Logic is the logic underlying modes of reasoning which are approximate. In fuzzy logic everything goes by membership function. A fuzzy condition is an assessment of a physical condition that is not measured with precision, but is assigned an intitutive method. In fact everything in the Universe is a little fuzzy and not digital, no matter how good measuring equipment is. A fuzzy set is a group of anything that cannot be precisely defined. For example, 'how hot the room is' the human might classify it at .2, if temperature is below freezing and might rate it at .9 or 1.0, if it is a hot day in summer. These perceptions are fuzzy which are not precisely measured facts.

Cluster analysis divides data into clusters or groups in the manner that similar data objects belong to the same cluster and dissimilar data objects belong to different clusters [7].

Knowledge discovery informs about techniques in the field of pattern recognition and databases [5, 18, 19]. In this process data mining is also a step for extracting models or patterns from given data [20, 21]. Pattern recognition is strengthened by Fuzzy Logic due to availability of more data [1, 2, 5, 22].

Partition clustering algorithms divide the data sets into cluster or classes whereas dissimilar data objects should belong to different clusters. In hierarchical clustering: objects that belong to a child cluster also belong to the parent cluster. In real life applications there is no sharp boundary between clusters so that fuzzy clustering is the only method suited for the data [3, 8, 11]. Instead of crisp assignments of the data to clusters, degree of member between zero and one is used in fuzzy clustering. Cluster analysis is a way of breaking data down into related components in such a way that patterns and order is recognized [5]. Cluster Analysis is a process of moving through large volumes of data in order to reveal useful pattern, useful information or clusters for any predefined analysis [4]. Clusters are grouping of data based on similarity metrics or probability density. As more and more data is available cluster analysis strengths the exposure of patterns [5, 20]. A fuzzy term membership is defined by measuring the distance from each cluster centers to the data point. The most prominent fuzzy clustering algorithms are Fuzzy C-means, Fuzzy K-means, (ISODATA), Gustafson-Kessel (GK) Algorithm. Fuzzy cluster analysis is used for the applications like Database, Pattern Recognition, Data Analysis, Detection of special geometrical shapes and image segmentation [5, 17, 20, 23]. This paper provides an overview of crisp clustering, advantages and limitations of fuzzy c-means clustering, comparison of fuzzy c-Means with Gustafson-Kessel fuzzy clustering algorithm, details of limitations of fuzzy c-means, defining basic notions of clustering and defining concept of fuzziness in a natural data [15].

2. FUZZY CLUSTERING TECHNIQUES
In fuzzy clustering techniques data is segmented and clusters are defined by grouping related attributes in uniquely defined clusters. A data point in the sample space is assigned to only one cluster and has an identity with it. When partioning is done in the data, the cluster centers are moved and not the data points [5]. In this clustering there is a self-iterative process of defining better cluster centers in each iteration. The most well known methods and commonly used portioning method is K-means algorithm. Here k denotes the number of cluster seeds initially provided for this algorithm. Here input parameter is taken as k and it partition set of m objects into k clusters

Here the formula 1 and technique is by computing the Euclidian distance between a data point and the cluster center to add item into one of the clusters resulting in high intra cluster similarity and low inter cluster similarity. The way to calculate the Euclidian distance is by calculating the sum of squared differences [9] and is defined as follows [10]:

$$d_k = \sum_n \left\| X_j^k - Ci_j \right\|^2 \qquad (1)$$

Where

d_k : is the distance of the k^{th} data point

n : is the number of attributes in a cluster

X_j^k : is j^{th} value of the k^{th} data point

C_j^i : is the j^{th} value of the i^{th} cluster centre

There is initialization on a random basis of the centres which are cluster *t* and a data point x_i to a cluster for which the distance is the least distance. After all these data points are assigned to various clusters, calculation is done for new cluster centres after calculating the weighted average for all data points in a cluster. This calculation of centre of clusters results in the movement towards the centre of cluster set. This process is repeated again and again till the change in cluster centres is nil.

The data in the physical world is never arranged in the defined very clear groups. Actually clusters will have not very clear boundaries and these not clear boundaries go into the space created by data in which very often there is overlapping regarding the boundaries of clusters. This is because in the physical world the data does not appear in a defined boundary but the data in the physical world has following drawbacks [6, 12, 13]:

- Unclear : In the physically world it is not definite
- Doubtful : It is not having all the information.
- Ambiguous: Many results can be derived from this.
- Vulnerable to change: The data in the physical world is vulnerable to change.
- Parameters which can change: These parameters which can change are not dependable.

After checking the above discussion, the two categories of limitations can be seen as vagueness and uncertainty [6]. In uncertainty the choice between two or more alternative is not identified. Through the use of fuzzy sets the knowledge of, not precise and concerned with quality, and not certain, is possible. These things can be done only through fuzzy logic as it unfolds akin to reasoning in the human specimen. It allows partial membership for data items which is not allowed in traditional logic but which is allowed in this concept.

3. FUZZY CLUSTERING ALGORITHMS

In fuzzy clustering, membership value is assigned to clusters. Clustering algorithms which are fuzzy allow clusters to grow. Membership value is very low in some cases indicating that the concerned data point may not be a member of the cluster under consideration. Some of the crisp techniques has difficulties in tackling outliers but in case of fuzzy techniques these outlier points are given small degree of membership. Degree to which the data point represents a cluster or not, is shown by the membership degree. Therefore fuzzy logic is the only method to handle data which is vague or which is not certain.

3.1 Fuzzy C-Means

Fuzzy C-Means (FCM) algorithm involves the processes in which there is calculation of cluster centers and assignment of points to these centres using a formula which is well known as Euclidian distance. The above process is kept on repeating itself until the stabilization of cluster centers. This algorithm assigns a membership value to the data items for the clusters within a range of 0 to 1. Thus the concepts of fuzzy sets of partial membership [14] are incorporated and forms overlapping clusters for supporting it. This algorithm needs a parameter which is called fuzzification parameter m which is in the range [1, *n*]. This fuzzification parameter determines the degree of fuzziness in these clusters. As the value of m equals one this very algorithm works as if it is a crisp portioning algorithm and the overlapping of clusters tend to be more for larger values of *m*.

The algorithm calculates the membership value μ with the following formula

$$\mu j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \qquad (2)$$

where,

$\mu_j(x_i)$: Is the membership of x_i in the j^{th} cluster

d_{ji} : is the distance of x_i in cluster c_j

m: is the fuzzification parameter

p: is the number of specified cluster

d_{ki} : is the distance of xi in cluster c_k

The new cluster centres are calculated with membership value using following equation

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \qquad (3)$$

Using equation 3 to calculate new membership values is done and is calculated as follows:

$$c_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m} \qquad (4)$$

where,

C_j : is the centre of the j^{th} cluster

x_i : is the i^{th} data point

μ_j : the function which returns the membership

m: is the fuzzification parameter

A form of weighted average is calculated using this special form. Degree of fuzziness in x_i 's current membership is calculated and this is multiplied by x_i . Sum of the

fuzzified membership divides the product obtained. New centroids are calculated in this manner.

Pseudo code is written as [16]:

Initialize p = number of cluster

Initialize m = fuzzification parameter

Initialize C_j *(cluster centre)*

Repeat

For i = 1 to n: Update $\mu_j(x_i)$ *applying (4)*

For j =1 to p: Update C_i *with (3) with current* $\mu_j(x_i)$

Firstly Centre of the cluster (normally called cluster centre) is calculated as weighted mean of the data. These weights are depending on the considered algorithm.

The matrix which is here the covariance matrix is defined as a fuzzy equivalence of classic covariance. From above equation, a constraint on the size is restricted on the covariance matrix whose determinant when calculated must be 1. Due to this, the Gustafson-Kessel clustering algorithm can identify ellipsoidal clusters having approximately the same size until C_j estimate stabilize

3.2. GUSTAFSON KESSEL FUZZY CLUSTERING ALGORITHM

The Gustafson-Kessel (GK) clustering algorithm is commonly used in most of the cases as a powerful clustering technique with various applications in various domains including image processing, classification and system identification. By contrast to the FCM algorithm its main feature is the local adaptation of the distance metric to the shape of the cluster. In addition it is not sensitive to the data scaling and initialization of the partition matrix.

Most of the clustering methods are based on the concept of batch clustering i.e. data set is assumed to be available before the clustering analysis is carried out. In a number of applications, however, data is presented to the clustering algorithm in real time and a growing number of methods try to cope the problem of evolving data stream clustering. Large amount of these methods are based on the single pass clustering comprising techniques that find clusters by executing one pass through the data set, in contrast to the iterative strategies like K-Means and FCM based clustering. Although the GK algorithm has a great advantage against the other clustering algorithms as it adapts the clusters according to the real shape of the cluster. The obstacles in this algorithm are due to an advance assumption of number of clusters and to determine clusters by a scheme which is iterative optimization.

However signifying f_{jk} which is the influence of point j on cluster k , the cluster centre and covariance matrix are calculated as follows:

$$w_k = \sum_{j=1}^{n} f_{jk}^{m} \; x_j$$

$$A_k = p\sqrt{\det\,(S_k)\; S_k^{-1}}$$

$$S_k = \sum_{j=1}^{n} f_{jk}^{m} \; (x_j - w_k)\,(x_j - w_k)^T$$

m is a parameter which has been defined by the user. This parameter is called fuzzifier. This step of updating the cluster parameter is alternated with the update of the coefficients which are weighing coefficients till a convergence criterion is met [9]. Now some choices for these weights are discussed. These are based on comparison between cluster centres and data. They rely on the distance which is calculated from the formula, $d_{jk} = (x_j - w_k)^T A_k^{-1} (x_i - w_k)$

4. EXPERIMENTAL ANALYSIS

The Experimental Analysis results in the Figure 1 and 2 show that Fuzzy C-Means have difficulty in handling outlier points.

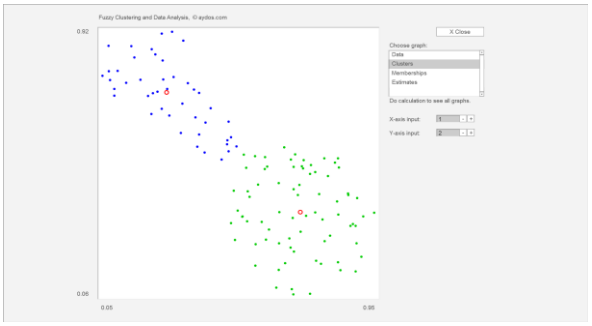

Fig 1. Graph between Fuzzy C-Means and Improved Fuzzy Functions

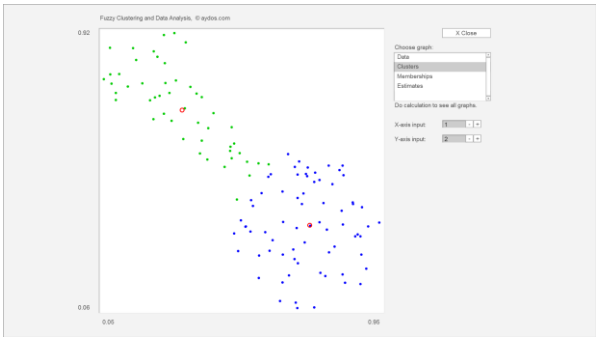

Fig 2. Graph between Gustafson-Kessel and Improved Fuzzy Functions

5. CONCLUSION

In the given algorithm Fuzzy c-means suffers from many constraints and these are due to the restriction that sum of membership value equal to one as is shown in the experimental analysis. Due to this drawback high membership value is given for the points which are on the outer of the clusters. Thus the algorithms have difficulty in handling outlier points. Also the membership of a data point in cluster depends directly on the other clusters and their membership values and due to these sometimes undesirable results is received from this algorithm and shows one of the major limitations of this algorithm

6. REFERENCES

[1] Kasabov, N. K. and Song, Q. 2002 "DENFIS: Dynamic Evolving Neural Fuzzy Inference system and its application for Time-Series Prediction" IEEE transactions on Fuzzy system, Vol. 10(2), pp. 144-154.

[2] Jian Yu, Miin-Shen Yang. 2007. "A Generalized Fuzzy Clustering Regularization Model with Optimally Tests and Model Complexity Analysis" IEEE transactions on Fuzzy System Vol. 15(5), pp. 904-915.

[3] Sato, M. and Sato, Y. 1995. "Fuzzy clustering model for fuzzy data", Fuzzy systems,1995, International Joint conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium Proceedings of 1995, IEEE International, Vol. 4, pp. 2123-2128.

[4] Bezdek, J. C. 1973. "Fuzzy Mathematics in Pattern classification" Ph.D. Thesis Centre for Applied Mathematics Cornell University, N.Y.

[5] Han, J., Kamber, M., 2006. "Data Mining: Concepts and Techniques, Second Edition" , Morgan Kaufmann.

[6] Klir G. J., Folger T. A., 1998. "Fuzzy sets, Uncertainty and information", Prentice Hall.

[7] Francisco de A.T. de Carvalho, Camilo P. Tenorio. 2010. "Fuzzy K-means clustering algorithms for interval valued data based on adaptive quadratic distances", Fuzzy Sets and Systems, Vol. 161(23), pp. 2978-2999.

[8] Chen, K.C.C, Au, W-H., Keith, Choi, B. 2002. "Mining Fuzzy rules in a Donor Database for Direct Mining by a charitable organization" Proceedings First IEEE international Conference on Cognitive Informatics, pp. 239-246.

[9] M.H. Fazel Zarandi, Zahara S. Razaee. 2010. "A Fuzzy Clustering Model for Fuzzy Data with Outliers", International Journal of Fuzzy System Applications, Vol. 1(2), IGI Global Publishers.

[10] Xiang Li, Hau-San Wong, Si. Wu. 2012. "A fuzzy minimax clustering model and its applications" Information Sciences: an International Journal, Vol. 186 (1), 114-125, Elsevier Science Inc..

[11] Pierpaolo D'Urso., Paolo Giordani. 2006. "A weighted fuzzy c-means clustering model for fuzzy data", Computational Statistics & Data Analysis Vol. 50 (6), pp. 1496-1523, Elsevier Science Pub..

[12] Inmon, W.H. 1996. "The data warehouse and data mining", Communications of ACM, Vol . 39 (11), pp. 49-50.

[13] M. Halkidi, D. Spinellis, G. Tsatsaronis, M. Vazirgiannis. 2011. "Data mining in software engineering", Intelligent Data Analysis Journal, Vol. 15(3).

[14] Au, W-H, Chan, K.C.C. 2001. "Classification with Degree of Membership: A Fuzzy Approach" Proceedings IEEE International Conference on Data mining, *(*ICDM 2001), pp. 35-42.

[15] Kruse, R., Borgelt, C, Nauck, D. 1999. "Fuzzy Data Analysis Challenges and Perspective", Fuzzy System Conference Proceedings, Vol. 3, pp. 1211-1216.

[16] Mitra, S., Pal, S.K., Mitra, P. 2002. "Data Mining in Soft Computing Framework : A Survey", IEEE transactions on neural networks, Vol. 13(1), pp. 3-14.

[17] http://fuzziness.org/fcm

[18] Kaur, H., Chauhan, R., Wasan, S. K. 2014. "A Bayesian Network Model for Probabilistic Estimation", Encyclopedia of Information Science and Technology, Third Edition, IGI Publishers, USA.

[19] Chauhan, R., Kaur, H. 2014. "Predictive Analytics and Data Mining: A framework for optimizing decisions with R tool", Advances in Secure Computing, Internet Services, and Applications, 73-88, IGI Publishers, USA.

[20] Kaur, H., Chauhan, R., Aljunid ,S. 2012. "Data Mining Cluster analysis on the influence of health factors in Casemix data", BMC Journal of Health Services Research, (June 2012). 12:O3

[21] Kaur, H., Chauhan, R., Alam, A. M. 2011. "Spatial Clustering Algorithm using R tree", Journal of Computing, 3(2), pp. 85-90, 2011.

[22] Malhotra, V.K., Kaur, H., Alam A. M. 2013. "A Spectrum of Fuzzy Clustering Algorithms and its applications", IEEE International Conference on Machine Intelligence, Research and Advancement (ICMIRA), pp. 599-603.

[23] Wasan, S. K., Bhatnagar, V., Kaur, H**.** 2007. "An Efficient Interestingness based Algorithm for Mining Association Rules in Medical Databases", Advances in Systems, Computing Sciences and Software Engineering, Springer, pp. 167-172.