2015 Fifth International Conference on Communication Systems and Network Technologies

# A Modified Fuzzy C Means Clustering using Neutrosophic Logic

Nadeem Akhtar
Department of Computer Engineering
Zakir Husain College of Engineering & Technology
Aligarh Muslim University, Aligarh, India
nadeemalakhtar@gmail.com

Mohd Vasim Ahamad
Department of Computer Engineering
Zakir Husain College of Engineering & Technology
Aligarh Muslim University, Aligarh, India
vasim.iu@gmail.com

*Abstract*— **A cluster can be defined as the collection of data objects grouped into the same group which are similar to each other whereas data objects which are different are grouped into different groups. The process of grouping a set objects into classes of similar objects is called clustering. In fuzzy c means clustering, every data point belongs to every cluster by some membership value. Hence, every cluster is a fuzzy set of all data points. Neutrosophic logic adds a new component "indeterminacy" to the fuzzy logic. In Neutrosophic Logic, the rule of thumb is that every idea has a certain degree of truthiness, falsity and indeterminacy which are to be considered independently from others. In our proposed algorithm, we have used Neutrosophic logic to add the indeterminacy factor in the Fuzzy C-Means Algorithm. We have modified the formula of calculating the membership value as well as the cluster center calculation and generated the clusters of documents as output.**

**Keywords- Clustering; Fuzzy Logic; Fuzzy C Means; Neutrosophic Logic**

## I. INTRODUCTION

In the information technology world, there is huge availability of data. But, this huge amount of data is of no use until it is converted into some useful information. Data mining can be defined as the process of extracting meaningful information from the huge amount of data available. Many technologies are used in data mining such as association rules discovery, clustering, classification, sequential pattern mining, etc. Among them, 'Clustering' is one of the most popular technologies. In traditional clustering approaches like partitioning approach, each data object belongs to only one cluster. There cannot be a common data point in any two clusters. Unlike the traditional partitioning methods, in fuzzy c means clustering, every data point belongs to every cluster by some membership value. Hence, every cluster is a fuzzy set of all the data points.

Jiang et al. [8] introduced a hybrid approach for clustering that is based on fuzzy c-means algorithm and immune single genetic. A novel clustering approach that is based on Dempster–Shafer (DS) theory of belief functions is known as evidential clustering [9]. There are as many as four evidential clustering algorithms have been proposed. The first one is, evidential clustering method (EVCLUS) [10].

EVCLUS can be used for both the metric and nonmetric data as it does not use any explicit geometrical model of the data [10]. Second, evidential fuzzy c-means algorithm (ECM) [11]. ECM is much closer with Fuzzy C-Means algorithm and its variants. ECM uses Euclidean metric as its distance measure and each cluster is represented by a prototype. Third one, relational evidential c-means algorithm (RECM) [12]. RECM is a modified version of the ECM algorithm [12]. The RECM optimization procedure is computationally much more efficient than the gradient based procedure [12]. Last one is constrained evidential c-means algorithm (CECM) [13]. CECM was introduced into the pairwise constraints in ECM [13].

Neutrosophy provids a new concept which consider an event or entity in set [5]. Smarandache introduced and added a new concept "Indeterminacy" to the fuzzy set. Thus, we can define the neutrosophic set as an ordered triple N= (T, I, F), where T represents the degree of truthiness, F is the degree of falsity, and I the level of indeterminacy [5]. In this scenario, T, I and F are also called as Neutrosphic components.

### A. Motivation

- To develop a clustering approach that considers the indeterminacy factor of documents towards cluster centers.
- It should produce high quality clusters
- It should be time efficient approach

## II. BACKGROUND

In this section, we firstly defined the clustering. Then, we explained the fuzzy c means algorithm and neutrosophic logic which are the main concern of our proposed algorithm.

978-1-4799-1797-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CSNT.2015.164

### A. Clustering

A cluster can be defined as the collection of data objects grouped into the same group which are similar to each other whereas data objects which are different are grouped into different groups. The process of grouping a set objects into classes of similar objects is called clustering. Clustering can be referred as unsupervised learning which do not rely on predefined classes unlike classification. There are be some remarkable clustering methods like as Partitioning Methods, Density Based Methods, Model Based Approaches, Soft computing Methods, etc.

### B. Fuzzy C Means Clustering

Fuzzy C-Means clustering algorithm is most popular method under the soft computing clustering approaches. In traditional clustering approaches like partitioning approach, each data object belongs to only one cluster. It discovers the disjoint clusters. There cannot be a common data point in any two clusters.

Unlike the traditional partitioning methods, in soft computing method, every data point belongs to every cluster by some membership value. Hence, every cluster is a fuzzy set of all data points. The fuzzy C means algorithm was developed by Dunn in 1973.It is modified and enhanced by Bezdek in 1981. The Fuzzy C-Means algorithm works such that a data object may belong to every cluster with some membership values between 0 and 1.

The fuzzy c means algorithm starts by assigning membership value to each of the data object with respect to each cluster center based on the distance between the cluster center and the data object [4]. Membership value will be more if the data object is near to a cluster center. The sum of membership values of each data object should be equal to one [4]. At each iteration, the membership grades and cluster centers are updated.

### C. Neutrosophic Logic

The fuzzy logic has been introduced to develop the systems that deal with approximate and uncertain ideas. In fuzzy logic we have only two components associated with each idea i.e. degree of truthiness and degree of falsity. We can represent it as a set F= {T, F}, where T represents the set of degree of truthiness and F represents the set of degree of falsity.

Florentin Smarandache proposed a new branch of philosophy called as Neutrosophic logic that deals with the origin, nature and scope of neutralities, their interactions with different ideational spectra [5]. Smarandache introduced and added a new concept "Indeterminacy" to the fuzzy set and modified as N= {(T, I, F): T, I, F [0, 1]}. Thus we can define the neutrosophic set as an ordered triple N= (T, I, F), where T represents the degree of truthiness, F is the degree of falsity, and I the level of indeterminacy. In this scenario, T, I and F are also called as Neutrosphic components. Hence, neutrosophic logic can also evaluate the degree of indeterminacy in addition to degree of truth and falsity.

## III. PROPOSED WORK

In our proposed algorithm, we have used Neutrosophic logic to add the indeterminacy factor in the Fuzzy C-Means Algorithm. We have modified the formula of calculating the membership value as well as the cluster center calculation and generated the clusters of documents.

Let us suppose the dataset D has N number of documents and every document has d dimensions. Let C be the number of clusters required, which must be decided in advance. The aim of Modified Fuzzy C-Means Clustering Algorithm using Neutrosophic Logic is to group the N documents into C clusters by using "indeterminacy" factor, where each and every document have some truth membership grade, a level of indeterminacy and some false membership values with respect to every cluster.

The level indeterminacy of each data object greatly depends on the clusters that are determinate and close to it. Hence, we considered the two closest clusters that are determinate and that have the biggest and the second biggest membership grades. Higher the truth membership grade of a document towards a cluster, greater the chances to be associated with that cluster.

### A. Algorithm Steps

Given a dataset D, choose appropriate number of clusters c such that 1<c<N, weighted factor m such that m>1 (generally taken m=2), and the termination tolerance ε such that ε>0.

1. Select random initial centres
2. Initialize partition matrix U= $[u_{ij}]$, U(0) and indeterminacy matrix I= $[I_{ij}]$, I(0)
3. At k-step, calculate the center vector $c^{(k)}$=c[j] with $u^{(k)}$

$$c_j = \frac{\sum_{i=1}^{n} (I_{ij} \cdot u_{ij})^m \cdot x_i}{\sum_{i=1}^{n} (I_{ij} \cdot u_{ij})^m}$$

4. Update $U^{(k)}$, $U^{(k+1)}$, $I^{(k)}$, $I^{(k+1)}$

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

$$I_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - \overline{c_{avg}}\|}\right)^{\frac{2}{m-1}}}$$

Where $\overline{C_{avg}} = \frac{c_{pi}+c_{qi}}{2}$, pi and qi are the cluster numbers with the biggest and second biggest value of T.

5. if $\|u^{(k+1)}-u^{(k)}\| < \varepsilon$ or the minimum J is achieved then STOP else goto (3)

The Objective Function J can be defined as

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{c}(u_{ij})^m . (\|x_i - c_j\|)^2$$

Our proposed algorithm, takes the input dataset and preprocess the documents in the dataset. Word Stemming, feature generation and feature selection is done on the documents. After that, we have a vector representing the documents. Then algorithm takes the random values for the cluster centers of the required number of cluster (to be decided in advance).

It initialize the truth membership grade vector as well as level of indeterminacy vector. Then, it groups the documents into the same clusters whose truth membership is high and indeterminacy is low. After that, algorithm iterates in updating the cluster centers by taking the mean of document's distances. It also updates the level of indeterminacy using our modified formula which considers the two closest clusters that are determinate and that have the biggest and the second biggest membership grades.

The proposed algorithm iterates until a minimum objective function is achieved or maximum number of iterations have been encountered. At the end, we have the clusters of documents as the output, which can be seen using a web browser like Chrome, Firefox or Internet Explorer etc. By clicking on the cluster numbers, we can see the documents that are grouped into that cluster. If we click on any document, it will show the content of that particular document.

## IV. Simulation And Results

### A. Dataset Description

The Modified Fuzzy C-Means Clustering Algorithm using Neutrosophic Logic is tested on two dataset. One is anonymous dataset and another is Mini Newsgroup dataset. The characteristics of the datasets are shown in table below. In DS 1, we have 10 groups having a total of 1000 documents. In each group, documents are unevenly distributed. Some of them are in higher number and some in lesser number. It is done to check the Precision variance of the documents. In DS 2, we have 10 groups having a total of 995 documents. In each group, documents are somewhat evenly distributed.

TABLE I: DATASET DESCRIPTION

| Datasets | documents# | Clusters # | Source |
|---|---|---|---|
| Anonymous | 500 | 5 | Downloaded from the internet |
| DS 1 | 1000 | 10 | https://archive.ics.uci.edu/ml/machine-learning-databases/20newsgroups-mld/ |
| DS 2 | 995 | 10 | https://archive.ics.uci.edu/ml/machine-learning-databases/20newsgroups-mld/ |

### B. Results for Anonymous Dataset

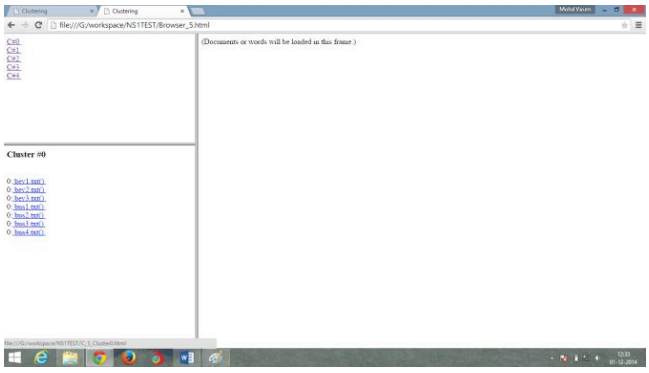


Figure 1. Representing cluster# 0 for anonymous dataset

### C. *Results for Dataset DS1*

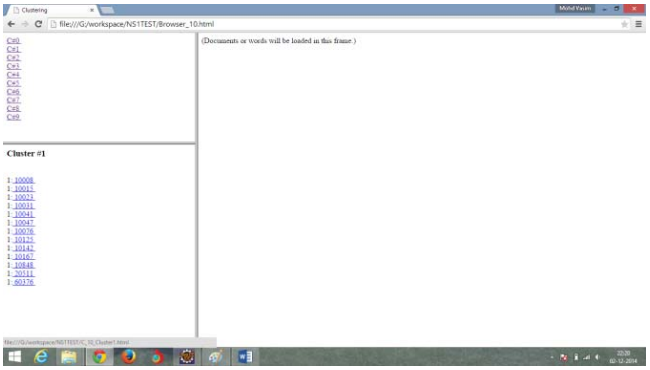

Figure 2. Representing cluster# 1 for dataset DS1

### D. *Results for Dataset DS2*

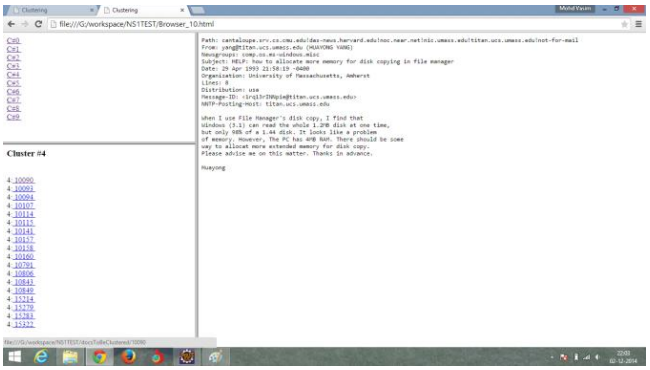

Figure 3. Representing cluster# 4 for dataset DS2

### E. *Performance Evaluation*

To measure the performance of our proposed algorithm, we have executed our algorithm on above two datasets and calculated the Precision and Recall based on the resultant clusters. We, also compared the results of our algorithm to the results of traditional Fuzzy C Means algorithm.
Precision and Recall can be calculated using the formula

Precision =TP / (TP+FN) %

Recall =TP / (FP+TN) %

Where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

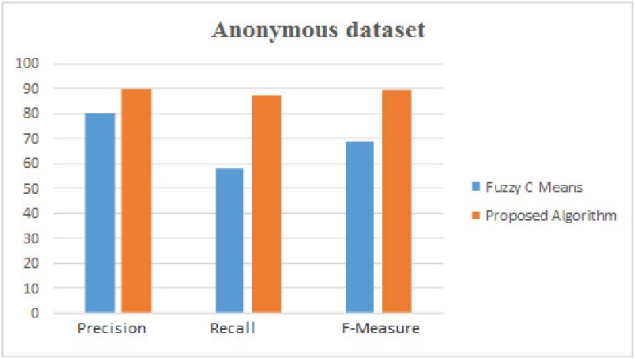


Figure 4. Comparison of FCM and Proposed algorithm for Anonymous dataset

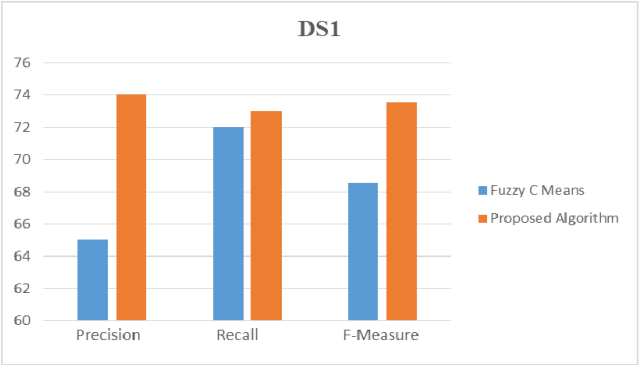


Figure 5. Comparison of FCM and Proposed algorithm for dataset DS1

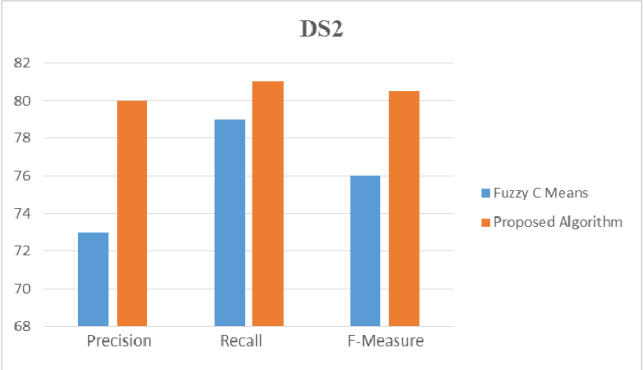


Figure 6. Comparison of FCM and Proposed algorithm for dataset DS2

## V. Conclusion and Future Work

The aim of Modified Fuzzy C-Means Clustering Algorithm using Neutrosophic Logic is to group the N documents into C clusters by using "indeterminacy" factor, where each and every document have some truth membership grade, a level of indeterminacy and some false membership values with respect to every cluster. The level indeterminacy of each data object greatly depends on the clusters that are determinate and close to it. Hence, we considered the two closest clusters that are determinate and that have the biggest and the second biggest membership grades. We have evaluated the performance of the proposed algorithm on the basis of the precision and recall. We also have compared the results with traditional FCM algorithm.

The proposed algorithm is sensitive to random initial centers affecting its efficiency. In the presence of random initial centers, if we execute the algorithm with same dataset many times, we will get different results every time. In future, we will try to mitigate this problem and propose the algorithm that will take optimized initial centers instead of random initial centers. There can be some improvements in cluster results also.

## References


[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer, Norwell, MA, 1981.

[3] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

[4] R.Suganya, R.Shanthi, "Fuzzy C- Means Algorithm- A Review" in International Journal of Scientific and Research Publications, Volume 2, Issue 11, November 2012.4.

[5] F. Smarandache. A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability. American Research Press, Rehoboth, NM, 1999.

[6] C. Ashbacher. Introduction to Neutrosophic Logic. American Research Press, Rehoboth, NM, 2002.

[7] F. Smarandache, R. Sunderraman, H. Wang, and Y. Zhang. Interval Neutrosophic Sets and Logic: Theory and Applications in Computing. HEXIS Neutrosophic Book Series, No. 5, Books on Demand, Ann Arbor, Michigan, 2005.

[8] Jiang H, Liu Y, Ye F, Xi H, Zhu M (2013) "Study of clustering algorithm based on fuzzy C means and immunological partheno genetic", J Softw 8(1):134–141.

[9] Smets P (1998) The transferable Belief Model for quantified belief representation. In: Gabbay DM, Smets P (eds) Handbook of defeasible reasoning and uncertainty management systems, vol 1. Kluwer Academic Publishers, Dordrecht, pp 267–301.

[10] DenWux T, Masson MH (2004) EVCLUS: evidential clustering of proximity data. IEEE Trans Syst Man Cybern Part B 34(1):95–109.

[11] Masson MH, Denoeux T (2008) ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recogn 41:1384–1397.

[12] Masson MH, Denoeux T (2009) RECM: relational evidential c-means algorithm. Pattern Recogn Lett 30:1015–1026.

[13] Antoine V, Quost B, Masson MH, Denoeux T (2012) CECM: constrained evidential c-means algorithm. Comput Stat Data Anal 56:894–914.

[14] Yanhui Guo • Abdulkadir Sengur, "NECM: Neutrosophic evidential c-means clustering algorithm" in Neural Comput & Applic DOI 10.1007/s00521-014-1648-3.

[15] Makhalova Elena, "FUZZY C - MEANS CLUSTERING IN MATLAB" in The 7th International Days of Statistics and Economics, Prague, September 19-21, 2013.

[16] K.Sathiyakumari, V.Preamsudha , G.Manimekalai, "Unsupervised Approach for Document Clustering Using Modified Fuzzy C mean Algorithm" in International Journal of Computer& Organization Trends – Volume1Issue3- 2011.

[17] M.S.Yang," A Survey of fuzzy clustering" Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.