International Journal of Computational Intelligence and Informatics, Vol. 3: No. 3, October - December 2013

Fuzzy Clustering using Credibilistic Critical Values

S. Sampath
Department of Statistics
University of Madras
Chennai, TamilNadu,India.
sampath1959@yahoo.com

R. Senthil Kumar
School of Advanced Sciences
VIT University Chennai
Chennai, TamilNadu,India.
statsenthil@gmail.com

***Abstract*- In this paper, the utility of credibilistic critical values in crisp conversion of fuzzy data sets is considered. Conversion of this type becomes essential mainly when clustering of fuzzy data sets is carried out. In this paper performance of two popular clustering algorithms namely Fuzzy *c*–means and Fuzzy *c*–medoids algorithms are evaluated under credibilistic critical value crisp conversion is carried out. Two synthetic data sets of varying nature are used in the comparative study. Some popular fuzzy clustering validity measures were employed in this study.**

Keywords- Clustering, Critical values, Credibility space, Partition Coefficient, Partition Entropy, FS Index, XB Index

I. INTRODUCTION

One of the major branches of data mining is cluster analysis. Cluster analysis finds applications in various branches of scientific studies which include image processing, mining of text data, mining of time series data, identification of patterns in spatial data, medical diagnostic studies etc. Ever since cluster analysis was introduced, researchers have developed various types of clustering algorithms of diversified nature. A vast majority of clustering algorithms available in literature are specifically meant for data of precise or crisp nature. The introduction of fuzzy set theory by [1] has led to databases consisting of imprecise data sets. Hence, there is a need for developing clustering algorithms specifically for imprecise or fuzzy data sets. The crux of the problem in dealing with imprecise data sets is the absence of proper definition of levels of similarity or dissimilarity between objects of the database which assume imprecise data values. There are several measures available for defining the dissimilarity between objects assuming imprecise values. Some of them are due to [2, 3, 4, 5, 6, 7, 8, 9]. Recently [10] and [11] have used certain measures of dissimilarity of imprecise data objects based on Credibility theory founded by [12]. They have studied the performance of *k*–means and *k*–medoids clustering algorithm by using such measures for certain data sets. In this paper, it is proposed to study the use of credibilistic critical values under Fuzzy *c*–means and Fuzzy *c*–medoids clustering algorithms. A comparative study on their performances in using the credibilistic critical values has been carried out with the help of few clustering validity measures which are specifically meant for fuzzy clustering algorithms. For detailed review on fuzzy clustering one can refer to the recent works of [13] and [14].

The paper is organized as follows: Section II provides the background of Credibility Theory, definition of fuzzy variable, Credibility distribution, Optimistic and Pessimistic values. Section III explains the process of creating crisp data set using the concept of credibilistic critical values. Section IV gives an overview of Fuzzy *c*–means clustering and Fuzzy *c*–medoids clustering along with some validity measures. Section V gives details of an experimental study carried out. Section VI presents conclusions drawn from this study.

II. CREDIBILITY THEORY

Credibility theory is a new branch of mathematics for studying the behavior of fuzzy phenomena introduced by Liu [2]. Some basic concepts and definitions related with credibility theory are stated below.

Let Θ be a non empty set and P be the power set of Θ. Each element in P is called an event. In order to present an axiomatic definition of credibility measure, it is necessary to assign a number $Cr(A)$ to each event A, where $Cr(A)$ indicates the credibility that A will occur. The following four axioms are satisfied by a credibility measure.

1. (Normality) $Cr(\Theta) = 1$
2. (Monotonicity) $Cr(A) \le Cr(B)$ whenever $A \subset B$
3. (Self Duality) $Cr(A) + Cr(A^c) = 1$
4. (Maximality) $Cr\{\bigcup_i A_i\} = Sup_i Cr\{A_i\}$ for any events $\{A_i\}$ with $Sup_i Cr\{A_i\} < 0.5$.

ISSN: 2349-6363

The triplet (Θ, P, Cr) is called a credibility space. A credibility measurable function from credibility space (Θ, P, Cr) to the set of real numbers is called a fuzzy variable. The membership function μ of a fuzzy variable ξ defined on the credibility space (Θ, P, Cr) is derived from the credibility measure by

$$\mu(x) = (2Cr\{\xi = x\}) \wedge 1, \forall x \in \Re \tag{1}$$

Membership function represents the degree that the fuzzy variable takes on some prescribed value. Here we shall provide some definitions related with credibility theory which are useful for understanding contents of this paper.

Credibility Distribution

The credibility distribution $\Phi : R \to [0,1]$ of a fuzzy variable ξ is defined by,

$$\Phi(x) = Cr\{\theta \in \Theta / \xi(\theta) \le x\} \tag{2}$$

Critical Values

For every fuzzy variable ξ, [12] defines two crisp critical values, namely, Optimistic and Pessimistic values as follows.

Optimistic Value

Let ξ be a fuzzy variable, and $\alpha \in (0,1]$. The α–optimistic value of ξ is defined as

$$\xi_{\sup}(\alpha) = \sup\{r \,|\, Cr\{\xi \ge r\} \ge \alpha\} \tag{3}$$

Pessimistic Value

Let ξ be a fuzzy variable, and $\alpha \in (0,1]$. The α–pessimistic value of ξ is defined as

$$\xi_{\inf}(\alpha) = \inf\{r \,|\, Cr\{\xi \le r\} \ge \alpha\} \tag{4}$$

It shows that the α–optimistic value $\xi_{\sup}(\alpha)$ is the supremum value that ξ achieves with credibility at least α, and the α–pessimistic value $\xi_{\inf}(\alpha)$ is the infimum value that ξ achieves with credibility at least α.

III. Crisp Conversion of Fuzzy Data Sets

Consider the data set consisting of *n* objects, namely $O_1, O_2, ..., O_n$ each having p attributes. In this study, we treat the values assumed by these objects with respect to these attributes as fuzzy variables having well defined credibility distributions. That is, the value assumed by the i^{th} object with respect to j^{th} attribute is treated as the fuzzy variable $\xi_{ij} (i = 1, 2, ..., n; j = 1, 2, ..., p)$. We shall assume that each of these fuzzy variables take m possible values $\xi_{ij}^u (u = 1, 2, ..., m)$. Under this set up the data matrix will appear as follows.

$$\begin{bmatrix} \xi_{11} = \begin{pmatrix} \xi_{11}^1 & \cdots & \xi_{11}^u & \cdots & \xi_{11}^m \\ \mu_{11}^1 & \cdots & \mu_{11}^u & \cdots & \mu_{11}^m \end{pmatrix} & \cdots & \xi_{1j} = \begin{pmatrix} \xi_{1j}^1 & \cdots & \xi_{1j}^u & \cdots & \xi_{1j}^m \\ \mu_{1j}^1 & \cdots & \mu_{1j}^u & \cdots & \mu_{1j}^m \end{pmatrix} & \cdots & \xi_{1p} = \begin{pmatrix} \xi_{1p}^1 & \cdots & \xi_{1p}^u & \cdots & \xi_{1p}^m \\ \mu_{1p}^1 & \cdots & \mu_{1p}^u & \cdots & \mu_{1p}^m \end{pmatrix} \\ \vdots & & \cdots & & \vdots & & \cdots & & \vdots \\ \xi_{i1} = \begin{pmatrix} \xi_{i1}^1 & \cdots & \xi_{i1}^u & \cdots & \xi_{i1}^m \\ \mu_{i1}^1 & \cdots & \mu_{i1}^u & \cdots & \mu_{i1}^m \end{pmatrix} & \cdots & \xi_{ij} = \begin{pmatrix} \xi_{ij}^1 & \cdots & \xi_{ij}^u & \cdots & \xi_{ij}^m \\ \mu_{ij}^1 & \cdots & \mu_{ij}^u & \cdots & \mu_{ij}^m \end{pmatrix} & \cdots & \xi_{ip} = \begin{pmatrix} \xi_{ip}^1 & \cdots & \xi_{ip}^u & \cdots & \xi_{ip}^m \\ \mu_{ip}^1 & \cdots & \mu_{ip}^u & \cdots & \mu_{ip}^m \end{pmatrix} \\ \vdots & & \cdots & & \vdots & & \cdots & & \vdots \\ \xi_{n1} = \begin{pmatrix} \xi_{n1}^1 & \cdots & \xi_{n1}^u & \cdots & \xi_{n1}^m \\ \mu_{n1}^1 & \cdots & \mu_{n1}^u & \cdots & \mu_{n1}^m \end{pmatrix} & \cdots & \xi_{nj} = \begin{pmatrix} \xi_{nj}^1 & \cdots & \xi_{nj}^u & \cdots & \xi_{nj}^m \\ \mu_{nj}^1 & \cdots & \mu_{nj}^u & \cdots & \mu_{nj}^m \end{pmatrix} & \cdots & \xi_{np} = \begin{pmatrix} \xi_{np}^1 & \cdots & \xi_{np}^u & \cdots & \xi_{np}^m \\ \mu_{np}^1 & \cdots & \mu_{np}^u & \cdots & \mu_{np}^m \end{pmatrix} \end{bmatrix}$$

When a data matrix of the above form is available for the process of a clustering, one faces the task of a defining the dissimilarity levels between different pairs of the objects in the data set. In order to make use of well defined dissimilarity measures meant for crisp data sets one can think of converting the above fuzzy data set into a crisp data set by using certain tools available in credibility theory. [10] used the concept of credibility critical values and [11] used the concept of credibility expectation in creating crisp data sets. These two approaches of creating crisp data sets have been compared using *k*–means and *k*–medoids clustering algorithms for certain synthetic data sets. It was found that the use of credibility critical values in creating crisp data sets is more efficient than using credibility expectation. Therefore, in this study we restrict ourselves to crisp conversion of the fuzzy data set using credibility critical values. It may be noted that for every fuzzy variable one can determine the pessimistic and optimistic values for a pre-determined value α on using for the definitions stated in the previous section. We shall denote these pessimistic and optimistic values by $\overline{\xi}_{ij}$ and $\underline{\xi}_{ij}$ respectively. Using these values one can develop another crisp value on averaging them. We shall denote them

by $\overset{*}{\xi}_{ij} = \frac{1}{2}\left[\overline{\xi}_{ij} + \underline{\xi}_{ij}\right]$. Thus on using these three values, namely pessimistic, optimistic and average of pessimistic and optimistic values one can define three *p*-component crisp vectors, namely, $\overline{\xi}_i = \left(\overline{\xi}_{i1}, \overline{\xi}_{i2}, ..., \overline{\xi}_{ip}\right), \underline{\xi}_i = \left(\underline{\xi}_{i1}, \underline{\xi}_{i2}, ..., \underline{\xi}_{ip}\right)$ and $\overset{*}{\xi}_i = \left(\overset{*}{\xi}_{i1}, \overset{*}{\xi}_{i2}, ..., \overset{*}{\xi}_{ip}\right)$ for every objects in the data set. Once these values identified, any appropriate distance metric can be used to measure the levels of dissimilarity. Thus, we have three different dissimilarity measures as given below.

1. Distance based on pessimistic value $\underline{d}_{ij} = \sqrt{\sum_{k=1}^{p}\left(\underline{\xi}_{ik} - \underline{\xi}_{kj}\right)^2}$
2. Distance based on optimistic value $\overline{d}_{ij} = \sqrt{\sum_{k=1}^{p}\left(\overline{\xi}_{ik} - \overline{\xi}_{kj}\right)^2}$
3. Distance based on average of optimistic and pessimistic values $\overset{*}{d}_{ij} = \sqrt{\sum_{k=1}^{p}\left(\overset{*}{\xi}_{ik} - \overset{*}{\xi}_{kj}\right)^2}$

These distances can be used as an input and a clustering algorithm can be implemented.

For example, consider the following matrix *A* which describes the distributions of fuzzy variables corresponding to 5 objects each having 2 attributes.

$$A = \begin{bmatrix} \begin{pmatrix} 48.0711 & 49.0711 & 50.0711 \\ 0.4314 & 0.1361 & 1 \end{pmatrix} & \begin{pmatrix} 50.7267 & 51.7267 & 52.7267 \\ 0.0760 & 1 & 0.4893 \end{pmatrix} \\ \begin{pmatrix} 42.9337 & 43.9337 & 44.9337 \\ 1 & 0.8693 & 0.6221 \end{pmatrix} & \begin{pmatrix} 46.4788 & 47.4788 & 48.4788 \\ 0.2399 & 1 & 0.3377 \end{pmatrix} \\ \begin{pmatrix} 45.2957 & 46.2957 & 47.2957 \\ 1 & 0.5797 & 0.3510 \end{pmatrix} & \begin{pmatrix} 48.0275 & 49.0275 & 50.0275 \\ 1 & 0.9027 & 0.9001 \end{pmatrix} \\ \begin{pmatrix} 43.6667 & 44.6667 & 45.6667 \\ 0.2638 & 0.5499 & 1 \end{pmatrix} & \begin{pmatrix} 49.4081 & 50.4081 & 51.4081 \\ 0.1839 & 0.9448 & 1 \end{pmatrix} \\ \begin{pmatrix} 45.5391 & 46.5391 & 47.5391 \\ 0.1455 & 1 & 0.4018 \end{pmatrix} & \begin{pmatrix} 49.7971 & 50.7971 & 51.7971 \\ 1 & 0.4909 & 0.1112 \end{pmatrix} \end{bmatrix}$$

$$B = \begin{bmatrix} 50.0711 & 52.7267 \\ 44.9337 & 47.4788 \\ 46.2957 & 50.0275 \\ 45.6667 & 51.4081 \\ 47.5391 & 50.7971 \end{bmatrix}, C = \begin{bmatrix} 48.0711 & 51.7267 \\ 42.9337 & 47.4788 \\ 45.2957 & 48.0275 \\ 44.6667 & 50.4081 \\ 46.5391 & 49.7971 \end{bmatrix} \text{ and } D = \begin{bmatrix} 49.0711 & 52.2267 \\ 43.9337 & 47.4788 \\ 45.7957 & 49.0275 \\ 45.1667 & 50.9081 \\ 47.0391 & 50.2971 \end{bmatrix}.$$

In the following section we study the performance of Fuzzy *c*–means clustering algorithm and Fuzzy *c*–medoids clustering algorithm with the help of crisp data sets developed in the above explained manner. To main the readability of the paper we present below brief descriptions of Fuzzy *c*–means clustering algorithm and Fuzzy *c*–medoids clustering algorithm. Definitions of certain validity measures are also given.

IV. Fuzzy Clustering

In *k*–means and *k*–medoids algorithm the data set is divided into *k* disjoint and exhaustive clusters, where each object belongs to exactly one cluster. In Fuzzy *C*–Means clustering objects can belong more than one cluster. The objects nearer to the cluster center are assigned with higher membership values and the objects far from the cluster center are assigned with lower membership values.

A. Fuzzy C–Means Clustering Algorithm

The Fuzzy *C*–Means (FCM) algorithm is an iterative clustering method introduced by [15] and improved by [16]. This method produces *c* clusters by minimizing the objective function

$$J(U, c_1, c_2, ...c_c) = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} d_{ij}^2 \tag{5}$$

where m is any real number which is greater than 1 ($1 \le m < \infty$), u_{ij} is the degree of membership of i^{th} object with cluster j, d_{ij} is the distance between the i^{th} object and the weighted centroid of the j^{th} cluster denoted by

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}, j = 1, 2, ..., c \quad (6)$$

The steps involved in Fuzzy c–means clustering algorithm are as follows.

Step 1: Fix the number of clusters c, where c is $(2 \le c < N)$ and the termination tolerance value ε.

Step 2: Select the parameter fuzziness exponent value m, where $1 \le m < \infty$.

Step 3: Calculate the fuzzy membership values by using the formula

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}} \right)^{\left(\frac{2}{m-1} \right)}}$$

Step 4: Calculate the c number of centroid vectors on using the formula

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}, j = 1, 2, ..., c$$

Step 5: Find the value of the objective function

$$J(U, c_1, c_2, ...c_c) = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m d_{ij}^2 .$$

Step 6: Repeat the step 3 to step 5 until $\max_{i,j} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^k \right| \right\} < \varepsilon$, for prespecified $\varepsilon > 0$.

B. Fuzzy c–medoids Clustering Algorithm

Fuzzy c–medoids (FCMdd) algorithm is an iterative clustering method introduced by [17]. Consider X be a set of N objects, $X = \{x_i | i = 1, 2, ..., N\}$. Let $d(x_i, x_j)$ be the distance between the object x_i and x_j. Let $V = \{v_1, v_2, ..., v_c\}, v_j \in X$ represent a subset of the object set X with c number of elements. Like Fuzzy c–means method, this method also produces c clusters by minimizing the objective function

$$J(V, X) = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m d(x_i, v_j) \quad (7)$$

where m is a fuzzy exponent or fuzzifier which is greater than 1 ($1 \le m < \infty$), u_{ij} is the degree of membership value of i^{th} object of cluster j, $d(x_i, v_j)$ is the distance or dissimilarity between the i^{th} object and the j^{th} cluster center v_j.

Fuzzy C–Medoids clustering method is carried out by optimizing the objective function in an iterative process. The steps involved in Fuzzy C–Medoids clustering algorithm are as follows.

Step 1: Fix the number of clusters *c*, where *c* is $(2 \le c < N)$ and the termination tolerance value ε.

Step 2: Select the parameter fuzziness exponent value *m*, where $1 \le m < \infty$.

Step 3: Randomly choose the initial set of medoids $V^{old} = \{v_1, v_2, ..., v_c\}$ from X_c.

Step 4: Calculate the fuzzy membership values by using the formula

$$u_{ij} = \frac{\left(\frac{1}{d(x_i, v_j)}\right)^{1/(m-1)}}{\sum_{k=1}^{c}\left(\frac{1}{d(x_i, v_k)}\right)^{\left(\frac{1}{m-1}\right)}}.$$

Step 5: Compute the new medoids V.

Step 6: Find the value of the objective function

$$J(V, X) = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} d(x_i, v_j)$$

Step 6: Repeat the step 3 to step 6 until $V^{old} = V$ or $\max_{i,j}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{k}\right|\right\} < \varepsilon$, for prespecified $\varepsilon > 0$.

C. Cluster Validity Measures

By applying a suitable clustering algorithm, one can partition the given data set. The clustering validity measures can help us to validate the partitioning by a numeric value. There are a number of validity measures available in order to validate fuzzy clustering algorithms. Some popular validity measures are Partition coefficient, Modified Partition Coefficient, Partition Entropy, Fukuyama and Sugeno Index and Xie and Beni Index. Definitions of these validity measures are presented below.

Partition Coefficient

Partition coefficient corresponding to a fuzzy partition of the data sets is defined as

$$PC = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{2} \tag{8}$$

where u_{ij} is the membership value of the i^{th} object with respect to j^{th} cluster. This index due to [18] takes values in the interval [1/c,1]. A value closer to 1 indicates clustering tends towards crisp clustering. It may be noted that a highly fuzzified partitioning is created when the membership values are closer to 1/c, which makes the value of partition coefficient very small. This indicates that a value closer to 1/c reveals the absence of a clustering tendency of objects in the given data sets.

Modified Partition Coefficient

To reduce the monotonic tendency of partition coefficient, an index proposed by [19] is defined as

$$MPC = 1 - \frac{c}{c-1}(1 - PC). \tag{9}$$

Partition Entropy

The partition entropy due to [20] is defined as

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij} \log_2 u_{ij}. \tag{10}$$

The Partition Entropy assumes values in the range [0,$\log_2$c]. The partition entropy will take the maximum value $\log_2 c$ and the minimum value zero. A value closer to 0 indicates that the clustering is crisper for the given data set and a higher value indicates the absence of clustering tendency of the objects.

Fukuyama and Sugeno Index (FS Index)

A validity measure proposed by [21] is

$$FS = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} \left\| x_i - v_j \right\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m} \left\| v_j - \bar{v} \right\|^2 = J_m(u,v) + K_m(u,v) \quad (11)$$

where $\bar{v} = \frac{1}{c}\sum_{j=1}^{c} v_j, v_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$. Here $J_m(u,v)$ is the objective function of the Fuzzy *c*–means clustering algorithm which measures the compactness of the clusters and $K_m(u,v)$ measures the separation. A smaller value of this index indicates a good partition.

Xie and Beni Index (XB Index)

A validity index defined by [22] and modified by [23] is

$$XB = \frac{\sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{2} \left\| x_i - v_j \right\|^2}{N \min_{i \neq j} \left\| v_i - v_j \right\|^2}. \quad (12)$$

A good clustering makes the value of *XB* index smaller.

In the following section of this paper we compare the performances of Fuzzy *c*–means and Fuzzy *c*–medoids clustering algorithms when crisp conversion of a fuzzy data set is carried out using credibilistic critical values. Two synthetic data sets generated from multivariate normal populations have been used in the comparative study.

V. Experimental Study

Dataset–1: The first data set consisting of 120 objects are generated from three different trivariate normal populations having equal mean vectors namely, [45 35 25] but with different variance–covariance matrices $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

The first variance-covariance matrix defines a system where the underlying variables are uncorrelated but with unequal variances, the second data set is associated with the variance-covariance matrix where all the components have unit variance and the third data set identifies the system where the variables have no constraints about their statistical dependency. Since the same mean vector is used for all the three populations, the data objects assume values which overlap.

From each of these three populations, 40 objects are generated using the *R* library function *mvrnorm*.

Dataset–2: The second data set containing 120 objects were generated in the above mentioned manner but with different mean vectors, namely, [45 35 25], [25 10 15] and [30 25 20]. Since the elements of these three mean vectors are separated to some extent. The data set generated in this manner is expected to have good separation in terms of the magnitude of values. We shall use the matrices constructed in the above manner to create fuzzy distributions by adopting the procedure described below.

Each element of the matrix which is nothing but a crisp quantity is used to create the credibility distribution of a fuzzy variable which can assume 5 different values. These values are obtained by adding the quantities 2, -1, 0, 1 and 2 with the given crisp value. Each of these values assigned a membership value in a random manner by generating 5 uniform random numbers from the interval [0, 1] ensuring that one of the values is 1. The membership value 1 is assigned to one of these 5 values in a random manner. For example, if 47.01 is a crisp number then the numbers generated are 45.01, 46.01, 47.01, 48.01 and 49.01. If the randomly generated uniform random numbers 0.0292, 0.6692, 1.0000, 0.6967, 0.5219 then the resulting fuzzy data is

$$\begin{array}{c} x \\ \mu(x) \end{array} \begin{bmatrix} 45.01 & 46.01 & 47.01 & 48.01 & 49.01 \\ 0.0292 & 0.6692 & 1.0000 & 0.6967 & 0.5219 \end{bmatrix}$$

Thus, the procedure of creating credibility distributions for the given crisp quantity is leads to the expansion of 120×3 data matrix into 120×15 data matrix. Assigning the membership value 1 to one of the possible values

of fuzzy variable is essential because a necessary and sufficient condition for a function to be a membership function is it should take value 1 for at least one *x*.

The enlarged data matrices were considered as fuzzy data sets. This fuzzy data set are converted into crisp data set by using the concept of pessimistic, optimistic and average values by taking three different levels of α, namely, 0.70, 0.80 and 0.90. Thus, we have generated nine sets of crisp data. Fuzzy *C*–Means and Fuzzy *C*–Medoids algorithms have been employed to these nine new data sets.

The data set consisting of 120 objects is partitioned into 3 clusters by using Fuzzy *C*–Means algorithm and Fuzzy *C*–Medoids algorithm. Since there are three classes in the data set, the number of clusters is taken as three. Clusters of objects have been formed using pessimistic, optimistic and average values.

Table I gives the values of various validity measures corresponding to different values of α used in the definition of critical values of a fuzzy variable for data set–1 (overlapping data set). It gives the values under both FCM and FCMdd clustering methods for all the three approaches of crisp conversion namely, pessimistic, optimistic and average.

TABLE I. CLUSTER VALIDITY MEASURES FOR THE DATA SET - 1

Pessimistic						
Critical Value	**0.70**		**0.80**		**0.90**	
Validity measures/Algorithm	**FCM**	**FCMdd**	**FCM**	**FCMdd**	**FCM**	**FCMdd**
Partition Coefficient	0.5653	0.6473	0.6053	0.6758	0.5664	0.6688
Modified Partition Coefficient	0.3479	0.4709	0.4079	0.5137	0.3496	0.5032
Partition Entropy	1.0746	0.8799	0.9846	0.8179	1.0660	0.8277
FS Index	383.2965	330.3434	306.8261	279.0392	352.3595	293.4524
Xie and Beni Index	0.9313	0.6103	0.7026	0.5667	0.9467	0.6298
Optimistic						
Critical Value	**0.70**		**0.80**		**0.90**	
Validity measures/Algorithm	**FCM**	**FCMdd**	**FCM**	**FCMdd**	**FCM**	**FCMdd**
Partition Coefficient	0.5403	0.6423	0.5810	0.6423	0.5653	0.6527
Modified Partition Coefficient	0.3104	0.4635	0.3714	0.4634	0.3479	0.4791
Partition Entropy	1.1295	0.8892	1.0393	0.8922	1.0734	0.8740
FS Index	404.4727	341.6970	315.0724	289.0516	314.0741	276.4682
Xie and Beni Index	1.1732	0.6007	0.7771	0.5695	1.0765	0.6417
Average of Pessimistic and Optimistic Values						
Critical Value	**0.70**		**0.80**		**0.90**	
Validity measures/Algorithm	**FCM**	**FCMdd**	**FCM**	**FCMdd**	**FCM**	**FCMdd**
Partition Coefficient	0.5413	0.6693	0.5635	0.6527	0.5723	0.6516
Modified Partition Coefficient	0.3120	0.5039	0.3453	0.4790	0.3585	0.4774
Partition Entropy	1.1255	0.8361	1.0764	0.8686	1.0538	0.8759
FS Index	323.0227	239.5323	290.8059	249.9105	308.7714	274.6766
Xie and Beni Index	0.9970	0.5440	0.8159	0.5342	0.9032	0.6752

Table II gives the values of various validity measures corresponding to different values of α used in the definition of critical values of a fuzzy variable for data set–2 (separated data set). It gives the values under both FCM and FCMdd clustering methods for all the three approaches of crisp conversion namely, pessimistic, optimistic and average.

TABLE II. CLUSTER VALIDITY MEASURES FOR THE DATA SET - 2

Pessimistic						
Critical Value	**0.70**		**0.80**		**0.90**	
Validity measures/Algorithm	**FCM**	**FCMdd**	**FCM**	**FCMdd**	**FCM**	**FCMdd**
Partition Coefficient	0.6222	0.6930	0.5581	0.6613	0.6450	0.6791
Modified Partition Coefficient	0.4333	0.5394	0.3372	0.4919	0.4675	0.5186
Partition Entropy	0.9397	0.7773	1.0862	0.8526	0.8864	0.7897
FS Index	313.3154	253.6830	275.9740	224.7249	241.6418	241.5876
Xie and Beni Index	0.4405	0.3817	0.8288	0.5481	0.3973	0.3184
Optimistic						
Critical Value	**0.70**		**0.80**		**0.90**	
Validity measures/Algorithm	**FCM**	**FCMdd**	**FCM**	**FCMdd**	**FCM**	**FCMdd**
Partition Coefficient	0.5924	0.6737	0.5619	0.6559	0.5939	0.6828
Modified Partition Coefficient	0.3886	0.5106	0.3429	0.4838	0.3908	0.5242
Partition Entropy	1.0119	0.8186	1.0733	0.8695	0.9956	0.7872
FS Index	399.109	349.7403	302.1060	254.4565	296.2683	224.7129
Xie and Beni Index	0.5928	0.4462	0.6459	0.5793	0.5662	0.3956
Average of Pessimistic and Optimistic Values						
Critical Value	**0.70**		**0.80**		**0.90**	
Validity measures/Algorithm	**FCM**	**FCMdd**	**FCM**	**FCMdd**	**FCM**	**FCMdd**
Partition Coefficient	0.6114	0.6959	0.5604	0.6748	0.6297	0.6820
Modified Partition Coefficient	0.4171	0.5438	0.3406	0.5121	0.4445	0.5230
Partition Entropy	0.9633	0.7707	1.0846	0.8209	0.9253	0.7940
FS Index	286.9409	234.5831	253.9687	178.4221	254.3573	216.6421
Xie and Beni Index	0.4447	0.4104	0.8019	0.4764	0.5454	0.3564

VI. CONCLUSIONS

This paper studies the choice of converting a fuzzy data set into a crisp data set using the concept of credibilistic critical values when fuzzy clusters are to be produced. It is to be noted that the quality of the fuzzy clusters depends on the choice of a clustering algorithm being used in this study. In this work the Fuzzy *c*–means and the Fuzzy *c*–medoids clustering algorithms have been evaluated with the help of some popular fuzzy clustering validity measures when crisp conversion is carried out with the different critical levels. The following are the observations made from the experimental study carried out in Section V.

i. In both the data sets FCMdd is shown perform well with respect to all the six clustering validity measures.

ii. Even though the superiority of a FCMdd is visible for all the data sets and all levels of crisp conversion, it is difficult to identify the best level used for crisp conversion.

iii. In the case of separated data sets crisp conversion using average of pessimistic and optimistic values with critical level 0.70 produces fuzzy clusters of a good quality.

iv. For overlapping data set pessimistic conversion with critical level 0.80 produces good quality clusters with respect to PC, MPC and PE.

REFERENCES

[1] L.A. Zadeh, "Fuzzy Sets," Information and Control , 8, 1965, pp. 338–353.

[2] C. Chakraborty and D. Chakraborty, " A fuzzy clustering methodology for linguistic opinions in group decision – making", Applied Soft Computing, 7 ,3, 2007, pp.858–869.

[3] C. H. Cheng and Y. Lin, "Evaluating the best main battle tank using fuzzy decision theory with linguistic criterion evaluation", Eur. J. Oper. Res., 142, 2002, pp.174–186.

[4] H.M. Hsu and C.T. Chen, "Aggregation of fuzzy opinions under group decision making", Fuzzy Sets Syst., 79, 1996, pp.279–285.

[5] H. S. Lee, "Optimal consensus of fuzzy opinions under group decision making environment", Fuzzy Sets Syst., 132, 2002, pp.303–315.

[6] J. Wang and Y. Lin, "A fuzzy multicriteria group decision making approach to select configuration items for software development", Fuzzy Sets Syst., 134, 2003, pp.343–363.

[7] R. C. Wang and S. J. Chuu,"Group decision-making using a fuzzy linguistic approach for evaluating the flexibility in a manufacturing system", Eur. J.Oper. Res., 154 (3), 2004, pp.563–572.

[8] S. Auephanwiriyakul and J. M. Keller, "Analysis and efficient implementation of a linguistic Fuzzy *c*–means", IEEE Trans. Fuzzy Syst., 10 (5), 2002, pp.563–582.

[9] V. F. Uricchio, R. Giordano and N. Lopez, "A fuzzy knowledge-based decision support system for groundwater pollution risk", J. Environ. Manage., 73 (3), 2004, pp.189–197.

[10] S. Sampath and R. KalaIvani, "Clustering of Fuzzy data using credibilistic critical values", In: IEEE International Conference on Signal and Image Processing, India, 2010, pp. 227–232.

[11] S. Sampath and R. Senthil Kumar, "Clustering of Fuzzy data using Credibilistic Expected and Critical Values", In: IEEE International Conference on Computer Communication and Systems, India,2014, pp.176–181.

[12] B. Liu, "Uncertainty Theory" ,http://orsc.edu.cn/liu/ut.pdf, 3rd ed., 2008.

[13] W. Pedrycz and G. Vukovichc, "Fuzzy clustering with supervision" ,Pattern Recognition, 37, 2004, pp.1339–1349.

[14] J. Valante de Oliveira and W. Pedrycz, "Advances in Fuzzy Clustering and its applications", John Wiley and Sons, New York, 2007.

[15] J.C. Dunn, "A fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Seperated Clusters", Journal of Cybernetics, 3 (3), 1973, pp.32–57.

[16] J.C. Bezdek, "Pattern recognition with fuzzy objective function algorithms", Plenum Press, New York, 1981.

[17] A. Joshi, L. Yi and R. Krishnapuram, "A Fuzzy Relative of the *k*–Medoids Algorithm with Application to Web Document and Snippet Clustering", in Fuzzy Systems, 1999.

[18] J.C. Bezdeck, R. Ehrlich and W. Full, "FCM: Fuzzy C – Means Algorithm", Computers and Geoscience, 1984.

[19] R. N. Dave, "Validating fuzzy partition obtained through c-shell clustering," Pattern Recognition letter 17, 1996, pp.613–623.

[20] J.C. Bezdek, "Cluster validity with fuzzy sets", Journal of Cybernet 3, 1974, pp.58–78.

[21] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method", In:Proc. 5th Fuzzy syst.Symp., 1989(in Japanese), pp.247–250.

[22] X. L. Xie and G. A. Beni, "Validity measure for fuzzy clustering", IEEE Transactions onPattern Analysis and Machine Intelligence , 3 (8),1991, pp.841–846.

[23] N.R. Pal and J.C. Bezdek, "On Cluster validity for the fuzzy c-means model", IEEE Trans. Syst., 3(3), 1995, pp.370–379.