An Enhanced Fuzzy Possibilistic C-means with Repulsion and Cluster Validity Index

D.Vanisri*
Department of Computer Technology,
Kongu Engineering College,
Perunudai-638 052, Tamilnadu,, INDIA

Dr.C.Loganathan
Principal, Maharaja Arts and Science College,
Coimbatore, Tamilnadu, INDIA

Abstract
The rapid worldwide increase in the data available leads to the difficulty for analyzing those data. Organizing data into interesting collection is one of the most basic forms of understanding and learning. Thus, a proper data mining approach is required to organize those data for better understanding. Clustering is one of the standard approaches in the field of data mining. The main of this approach is to organize a dataset into a set of clusters, which consists of "similar" data items, as calculated by some distance function. There are various clustering techniques like K-Means, Possibilistic C-Mean, etc., proposed by various researchers. Recently, Fuzzy Possibilistic C-Means is found to be better because of its embedded fuzzy logic. This paper initially proposed a Modified Fuzzy Possibilistic C-Means (MFPCM) algorithm which enhances the clustering accuracy. Next, Penalized and Compensated constraints are used in the objective function. For further improvement in clustering accuracy, Repulsion term is introduced in the objective function. Finally, Cluster Validity Index is performed by using Partition Coefficient and Exponential Separation (PCAES) method. The experimental result shows that the proposed clustering technique results in lesser error rate which in turn shows the better accuracy of classification.

Keywords
Fuzzy Possibilistic C-Means, Modified Fuzzy Possibilistic C-Means, Penalized and Compensated Constraints, Repulsion, Partition Coefficient and Exponential Separation

I. INTRODUCTION

THE growth and development in sensing and storage technology and drastic development in the applications such as internet search, digital imaging, and video surveillance have generated many high-volume, high-dimensional data sets. As the majority of the data are stored digitally in electronic media, they offer high prospective for the development of automatic data analysis, classification, and retrieval approaches. Clustering is one of the most popular approaches used for data analysis and classification. Fuzzy C-Means (FCM) algorithm [8] is a typical clustering algorithm which has been widely utilized in engineering and scientific disciplines such as medicine imaging, bioinformatics, pattern recognition, and data mining. As the basic FCM clustering approach employs the squared-norm to measure similarity between prototypes and data points, it can be effective in clustering only the 'spherical' clusters and many algorithms are derived from the FCM to cluster more general dataset. FCM approach is very sensitive to noise. To avoid such an effect, Krishnapuram and Keller removed the constraint of memberships in FCM and propose the Possibilistic C-Means (PCM) algorithm [15]. To classify a data point Pal deducted an approach that the data point must closely have their cluster centroid, and it is the role of membership. Also for the centroid estimation, the typicality is used for alleviating the unwanted effect of outliers. So Pal proposed a clustering algorithm called Fuzzy Possibilistic C-Means (FPCM) that combines the characteristics of both fuzzy [8] and possibilistic c-means [9]–[14]. In order to enhance the FPCM, Modified Fuzzy Possibilistic C-Means (MFPCM) approach is presented. This new approach provides better results compared to the previous algorithms by modifying the Objective function used in FPCM. The objective function is enhanced by adding new weight of data points in relation to every cluster and modifying the exponent of the distance between a point and a class.

The existing approach use the probabilistic constraint to enable the memberships of a training sample across clusters that sum up to 1, which means the different grades of a training sample are shared by distinct clusters, but not as degrees of typicality. In contrast, each component created by FPCM belongs to a dense region in the data set. Each cluster is independent of the other clusters in the FPCM [7] strategy. Typicalities and Memberships are very important factors for the correct feature of data substructure in clustering problem. If a training sample has been effectively

Manuscript received February 5, 2011
Manuscript revised February 20, 2011

classified to a particular suitable cluster, then membership is considered as a better constraint for which the training sample is closest to this cluster. In other words, typicality is an important factor to overcome the undesirable effects of outliers to compute the cluster centers. In order to enhance the above mentioned existing approach in MFPCM, penalized and compensated constraints are incorporated. Yang [16] and Yang and Su [17] have added the penalized term into fuzzy c-means to construct the penalized fuzzy c-means (PFCM) algorithm. The compensated constraint is embedded into FCM by Lin [18] to create compensated fuzzy c-means (CFCM) algorithm. In this paper the penalized and compensated constraints are combined with the MFPCM which is said to be Penalized and Compensated constraints based Modified Fuzzy Possibilistic C-Means clustering algorithm (PCMFPCM). Then the Repulsion [20] factors are embedded in the objective function to decrease the intercluster distance and at the same time increases the intracluster distance. Finally, Partition Coefficient and Exponential Separation (PCAES) [21] technique is used for cluster validity index.

II. RELATED WORKS

Clustering is found to be the widely used approach in most of the data mining systems. Compared with the clustering algorithms, the Fuzzy c means approach is found to be efficient and this section discusses some the literature studies on the fuzzy probabilistic c means approach for the clustering problem.

Pal et al., [1] proposed the Fuzzy-Possibilistic C-Means (FPCM) algorithm that generated both membership and typicality values when clustering unlabeled data. The typicality values are constrained by FPCM so that the sum of the over all data points of typicalities to a cluster is one. For large data sets the row sum constraint produces unrealistic typicality values. In this paper, a novel approach is presented called possibilistic-fuzzy c-means (PFCM) model. PFCM produces memberships and possibilities concurrently, along with the usual point prototypes or cluster centers for each cluster. PFCM is a hybridization of fuzzy c-means (FCM) and possibilistic c-means (PCM) that often avoids various problems of PCM, FCM and FPCM. The noise sensitivity defect of FCM is resolved in PFCM, overcomes the problem of coincident clusters of PCM and purges the row sum constraints of FPCM. The first-order essential conditions for extrema of the PFCM objective function is driven, and used them as the basis for a standard alternating optimization approach to finding local minima of the PFCM objective functional. With Some numerical examples FCM and PCM are compared to PFCM in [1]. The examples illustrate that PFCM compares favorably to both of the previous models. Since PFCM prototypes are fewer sensitive to outliers and can avoid coincident clusters, PFCM is a strong candidate for fuzzy rule-based system identification.

Xiao-Hong et al., [3] presented a novel approach on Possibilistic Fuzzy C-Means Clustering Model Using Kernel Methods. The author insisted that fuzzy clustering method is based on kernel methods. This technique is said to be kernel possibilistic fuzzy c-means model (KPFCM). KPFCM is an improvement in Possibilistic Fuzzy C-Means (PFCM) model which is superior to fuzzy c-means (FCM) model. The KPFCM model is different from PFCM and FCM which are based on Euclidean distance. The KPFCM model is based on non-Euclidean distance by using kernel methods. In addition, with kernel methods the input data can be mapped implicitly into a high-dimensional feature space where the nonlinear pattern now appears linear. KPFCM can deal with noises or outliers better than PFCM. The KPFCM model is interesting and provides good solution. The experimental results show better performance of KPFCM.

Ojeda-Magafia et al., [4] proposed a new technique to use the Gustafson-Kessel (GK) algorithm within the PFCM (Possibilistic Fuzzy c-Means), such that the cluster distributions have a better adaptation with the natural distribution of the data. The PFCM, proposed by Pal et al. on 2005, introduced the fuzzy membership degrees of the FCM and the typicality values of the PCM. However, this algorithm uses the Euclidian distance which gives circular clusters. So, combining the GK algorithm and the Mahalanobis measure for the calculus of the distance, there is the possibility to get ellipsoidal forms as well, allowing a better representation of the clusters.

Chunhui et al., [6] presented a similarity based fuzzy and possibilistic c-means algorithm called SFPCM. It is derived from original fuzzy and possibilistic-means algorithm (FPCM) which was proposed by Bezdek. The difference between the two algorithms is that the proposed SFPCM algorithm processes relational data, and the original FPCM algorithm processes propositional data. Experiments are performed on 22 data sets from the UCI repository to compare SFPCM with FPCM. The results show that these two algorithms can generate similar results on the same data sets. SFPCM performs a little better than FPCM in the sense of classification accuracy, and it also converges more quickly than FPCM on these data sets.

Yang et al., [5] puts forth an unlabeled data clustering method using a possibilistic fuzzy c-means (PFCM). PFCM is the combination of possibilistic c-means (PCM) and fuzzy c-means (FCM), therefore it has been shown that

PFCM is able to solve the noise sensitivity issue in FCM, and at the same time it helps to avoid coincident clusters problem in PCM with some numerical examples in low-dimensional data sets. Further evaluation of PFCM for high-dimensional data is conducted in this paper and presented a revised version of PFCM called Hyperspherical PFCM (HPFCM). The original PFCM objective function is modified, so that cosine similarity measure could be incorporated in the approach. When compared their performance with some of the traditional and recent clustering algorithms for automatic document categorization the FPCM performs better. The study shows HPFCM is promising for handling complex high dimensional data sets and achieves more stable performance. The remaining problem of PFCM approach is also discussed in this research.

A robust interval type-2 possibilistic C-means (IT2PCM) clustering algorithm is presented by Long Yu et al., [6] which is essentially alternating cluster estimation, but membership functions are selected with interval type-2 fuzzy sets by the users. The cluster prototypes are computed by type reduction combined with defuzzification; consequently they could be directly extracted to generate interval type-2 fuzzy rules that can be used to obtain a first approximation to the interval type-2 fuzzy logic system (IT2FLS). The IT2PCM clustering algorithm is robust to uncertain inliers and outliers, at the same time provides a good initial structure of IT2FLS for further tuning in a subsequent process. The better simulation results are obtained for the problem of classification and forecasting.

Sreenivasarao et al., [2] presented a Comparative Analysis of Fuzzy C- Mean and Modified Fuzzy Possibilistic C -Mean Algorithms in Data Mining. There are various algorithms used to solve the problem of data mining. FCM (Fuzzy C mean) clustering algorithm and MFPCM (Modified Fuzzy Possibililstic C mean) clustering algorithm are comparatively studied. The performance of Fuzzy C mean (FCM) clustering algorithm is analyzed and compared it with Modified Fuzzy possibilistic C mean algorithm. Complexity of FCM and MFPCM are measured for different data sets. FCM clustering technique is separated from Modified Fuzzy Possibililstic C mean and that employs Possibililstic partitioning. The FCM employs fuzzy portioning such that a point can belong to all groups with different membership grades between 0 and 1. The author concludes that the Fuzzy clustering, which constitute the oldest component of soft computing. This method of clustering is suitable for handling the issues related to understandability of patterns; incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. The proposed approach for the unlabeled data clustering is presented in the following section.

III. METHODOLOGY

3.1. Fuzzy Possibilistic Clustering Algorithm

The fuzzified version of the k-means algorithm is Fuzzy C-Means (FCM). It is a clustering approach which allows one piece of data to correspond to two or more clusters. Dunn in 1973 developed this technique and it was modified by Bezdek in 1981 [8] and this is widely used in pattern recognition. The algorithm is an iterative clustering approach that brings out an optimal c partition by minimizing the weighted within group sum of squared error objective function J_{FCM}:

$$J_{FCM}(V,U,X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d^2(X_j,v_i), 1 < m < +\infty \quad (1)$$

In the equation $X = \{x_1, x_2,...,x_n\} \subseteq R^p$ is the data set in the p-dimensional vector space, the number of data items is represented as p, c represents the number of clusters with $2 \leq c \leq n-1$. $V = \{v_1, v_2, \ldots, v_c\}$ is the c centers or prototypes of the clusters, v_i represents the p-dimension center of the cluster i, and $d^2(x_j, v_i)$ represents a distance measure between object x_j and cluster centre v_i. $U = \{\mu ij\}$ represents a fuzzy partition matrix with $u_{ij} = u_i(x_j)$ is the degree of membership of x_j in the ith cluster; x_j is the jth of p-dimensional measured data. The fuzzy partition matrix satisfies:

$$0 < \sum_{j=1}^{n} \mu_{ij} < n, \forall i \in \{1,...,c\} \quad (2)$$

$$\sum_{i=1}^{c} \mu_{ij} = 1, \forall j \in \{1,...,n\} \quad (3)$$

m is a weighting exponent parameter on each fuzzy membership and establishes the amount of fuzziness of the resulting classification; it is a fixed number greater than one. Under the constraint of U the objective function J_{FCM} can be minimized. Specifically, taking of J_{FCM} with respect to u_{ij} and v_i and zeroing them respectively, is necessary but not sufficient conditions for J_{FCM} to be at its local extrema will be as the following:

$$\mu_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d(X_j v_i)}{d(X_j v_k)}\right)^{2/(m-1)}\right]^{-1}, 1 \leq i \leq c, 1 \leq j \leq \quad (4)$$

$$v_i = \frac{\sum_{k=1}^{n} \mu_{ik}^{m} x_k}{\sum_{k=1}^{n} \mu_{ik}^{m}}, 1 \le i \le c. \quad (5)$$

In noisy environment, the memberships of FCM do not always correspond well to the degree of belonging of the data, and may be inaccurate. This is mainly because the real data unavoidably involves some noises. To recover this weakness of FCM, the constrained condition (3) of the fuzzy c-partition is not taken into account to obtain a possibilistic type of membership function and PCM for unsupervised clustering is proposed. The component generated by the PCM belongs to a dense region in the data set; each cluster is independent of the other clusters in the PCM strategy. The following formulation is the objective function of the PCM.

$$J_{PCM}(V,U,X) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ik}^{m} d^2(X_j, v_i) + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 \quad (6)$$

Where

$$\eta_i = \frac{\sum_{j=1}^{n} \mu_{ik}^{m} ||x_j - v_i||^2}{\sum_{j=1}^{n} \mu_{ij}^{m}} \quad (7)$$

η_i is the scale parameter at the ith cluster,

$$u_{ij} = \frac{1}{1 + \left[\frac{d^2(x_j, v_i)}{\eta_i}\right]^{\frac{1}{m-1}}} \quad (8)$$

u_{ij} represents the possibilistic typicality value of training sample x_j belong to the cluster i. $m \in [1,\infty]$ is a weighting factor said to be the possibilistic parameter. PCM is also based on initialization typical of other cluster approaches. The clusters do not have a lot of mobility in PCM techniques, as each data point is classified as only one cluster at a time rather than all the clusters simultaneously. Consequently, a suitable initialization is necessary for the algorithms to converge to nearly global minimum.

The characteristics of both fuzzy and possibilistic c-means approaches is incorporated. Memberships and typicalities are very important factors for the correct feature of data substructure in clustering problem. Consequently, an objective function in the FPCM depending on both memberships and typicalities can be represented as below:

$$J_{FPCM}(U,T,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} (\mu_{ij}^{m} + t^{m}) d^2(X_j, v_i) \quad (9)$$

with the following constraints :

$$\sum_{i=1}^{c} \mu_{ij} = 1, \forall j \in \{1, \ldots, n\} \quad (3)$$

$$\sum_{j=1}^{n} t_{ij} = 1, \forall i \in \{1, \ldots, c\} \quad (10)$$

A solution of the objective function can be obtained through an iterative process where the degrees of membership, typicality and the cluster centers are update with the equations as follows.

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d(X_j, v_i)}{d(X_j, v_k)}\right)^{2/(m-1)}\right]^{-1}, 1 \le i \le c, 1 \le j : \quad (4)$$

$$t_{ij} = \left[\sum_{k=1}^{n} \left(\frac{d(X_j, v_i)}{d(X_j, v_k)}\right)^{2/(\eta-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \quad (11)$$

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik}^{m} + t_{ik}^{\eta}) X_K}{\sum_{k=1}^{n} (\mu_{ik}^{m} + t_{ik}^{\eta})}, 1 \le i \le c. \quad (12)$$

PFCM constructs memberships and possibilities simultaneously, along with the usual point prototypes or cluster centers for each cluster. Hybridization of possibilistic c-means (PCM) and fuzzy c-means (FCM) is the PFCM that often avoids various problems of PCM, FCM and FPCM. The noise sensitivity defect of FCM is solved by PFCM, which overcomes the coincident clusters problem of PCM. But the estimation of centroids is influenced by the noise data.

3.2. Modified Fuzzy Possibilistic C-Means Technique (FPCM)

Objective function is very much necessary to enhance the quality of the clustering results. Wen-Liang Hung presented a new approach called Modified Suppressed Fuzzy c-means (MS-FCM), which significantly improves the performance of FCM due to a prototype-driven learning of parameter α [19]. Exponential separation strength between clusters is the

base for the learning process of α and is updated at each of the iteration. The parameter α can be computed as

$$\alpha = exp\left[-\min_{i \neq k}\frac{||v_i - v_k||^2}{\beta}\right] \quad (13)$$

In the above equation β is a normalized term so that β is chosen as a sample variance. That is, β is defined:

$$\beta = \frac{\sum_{j=1}^{n}||x_j - \bar{x}||^2}{n} \; where \; \bar{x} = \frac{\sum_{j=n}^{n} x_j}{n}$$

But the remark which must be pointed out here is the common value used for this parameter by all the data at each of the iteration, which may induce in error. A new parameter is added with this which suppresses this common value of α and replaces it by a new parameter like a weight to each vector. Or every point of the data set possesses a weight in relation to every cluster. Consequently this weight permits to have a better classification especially in the case of noise data. The following equation is used to calculate the weight.

$$w_{ji} = exp\left[-\frac{\|x_j - v_i\|^2}{\left[\sum_{j=1}^{n}\|x_j - \bar{v}\|^2\right] * c/n}\right] \quad (14)$$

In the previous equation w_{ji} represents weight of the point j in relation to the class i. In order to alter the fuzzy and typical partition, this weight is used. The objective function is composed of two expressions: the first is the fuzzy function and uses a fuzziness weighting exponent, the second is possibililstic function and uses a typical weighting exponent; but the two coefficients in the objective function are only used as exhibitor of membership and typicality. A new relation, lightly different, enabling a more rapid decrease in the function and increase in the membership and the typicality when they tend toward 1 and decrease this degree when they tend toward 0. This relation is to add Weighting exponent as exhibitor of distance in the two under objective functions. The objective function of the MFPCM can be given as follows:

$$J_{MFPCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m} w_{ij}^{m} d^{2m}(x_j, v) + t_{ij}^{\eta}\, w_{ij}^{\eta} d^{2\eta}(x_j \quad (15)$$

U = {μij} represents a fuzzy partition matrix, is defined as:

$$\mu_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d(X_j, v_i)}{d(X_j, v_k)}\right)^{2m/(m-1)}\right]^{-1} \quad (16)$$

T = {tij} represents a typical partition matrix, is defined as:

$$t_{ij} = \left[\sum_{k=1}^{n}\left(\frac{d(X_j, v_i)}{d(X_j, v_k)}\right)^{2\eta/(\eta-1)}\right]^{-1} \quad (17)$$

V = {vi} represents c centers of the clusters, is defined as:

$$v_i = \frac{\sum_{j=1}^{n}(\mu_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta}) * X_j}{\sum_{j=1}^{n}(\mu_{ik}^{m} w_{ji}^{m} + t_{ik}^{\eta} w_{ij}^{\eta})} \quad (18)$$

3.3. Penalized and Compensated constraints based Modified Fuzzy Possibilistic C-Means(PCMFPCM)

The Penalized and compensated constraints are embedded with the previously discussed Modified Fuzzy Possibilistic C-Means algorithm. The objective function of the FPCM is given in equation (15). In the proposed approach the penalized and compensated terms are added to the objective function of FPCM to construct the objective function of PCMFPCM. The penalized constraint can be represented as follows

$$\frac{1}{2} v \sum_{x=1}^{n}\sum_{i-1}^{c}(\mu_{xi}^{m} ln\alpha_i + t_{xi}^{\eta} ln\beta_x) \quad (19)$$

Where

$$\alpha_i = \frac{\sum_{x=1}^{n}\mu_{xi}^{m}}{\sum_{x=1}^{n}\sum_{i=1}^{c}\mu_{xi}^{m}}, \quad i = 1,2,\ldots\ldots c,$$

$$\beta_x = \frac{\sum_{i=1}^{c} t_{xi}^{\eta}}{\sum_{x=1}^{n}\sum_{i=1}^{c} t_{xi}^{\eta}} \quad x = 1,2,\ldots,n$$

where α_i is a proportional constant of class i; β_x is a proportional constant of training vector z_x, and v (v≥0); τ (τ≥0) are also constants. In these functions, α_i and β_x are defined in equations above. Membership $\mu_{xi}$and typicality t_{xi} for the penalize is presented below.

$$(\mu_{xi})_p = \left(\sum_{i=1}^{c}\frac{(\|z_x - \varpi_i\|^2 - v\, ln\alpha_i)^{1/(m-1)}}{(\|z_x - \varpi_i\|^2 - v\, ln\alpha_i)^{1/(m-1)}}\right)^{-1}$$

$$x = 1,2,\ldots,n, i = 1,2,\ldots c,$$

$$(t_{xi})_p = \left(\sum_{y=1}^{n}\frac{(\|z_x - \varpi_i\|^2 - v\, ln\beta_x)^{1/(\eta-1)}}{(\|z_x - \varpi_i\|^2 - v\, ln\beta_y)^{1/(\eta-1)}}\right)^{-1}$$

$x = 1,2, \dots, n, i = 1,2, \dots c,$

In the previous expression $\varpi_i = v_i = \frac{\sum_{k=1}^{n}(\mu_{ik}^{m}+t_{ik}^{\eta})x_k}{\sum_{k=1}^{n}(\mu_{ik}^{m}+t_{ik}^{\eta})}, 1 \le i \le c,$ which is the centroid. The compensated constraints can represented as follows

$$\frac{1}{2}\tau\sum_{x=1}^{n}\sum_{i=1}^{c}(\mu_{x,i}^{m}\tanh\alpha_i + t_{x,i}^{\eta}\tanh\beta_x) \quad (20)$$

Where Membership $\mu_{x,i}$ and typicality $t_{x,i}$ for the compenzation is presented below

$$(\mu_{x,i})_c = \left(\sum_{i=1}^{c}\frac{(\|z_x-\varpi_i\|^2 - \tau\tanh(\alpha_i))^{1/(m-1)}}{(\|z_x-\varpi_i\|^2 - \tau\tanh(\alpha_i))^{1/(m-1)}}\right)^{-1}$$

$x = 1,2, \dots, n, i = 1,2, \dots c,$

$$(t_{x,i})_c = \left(\sum_{y=1}^{n}\frac{(\|z_x-\varpi_i\|^2 - \tau\tanh(\beta_x))^{1/(\eta-1)}}{(\|z_x-\varpi_i\|^2 - \tau\tanh(\beta_y))^{1/(\eta-1)}}\right)^{-1}$$

$x = 1,2, \dots, n, i = 1,2, \dots c,$

To obtain an efficient clustering the penalization term must be removed and the compensation term must be added to the basic objective function of the existing FPCM. This brings out the objective function of PCFPCM and it is given in equation (21).

$$J_{MFPCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m}w_{ji}^{m}d^{2m}(x_j v) + t_{ij}^{\eta}w_{ji}^{\eta}d^{2\eta}(x_j,v_i)) - \frac{1}{2}v\sum_{x=1}^{n}\sum_{i=1}^{c}\left(\mu_{x,i}^{m}\ln\alpha_i + t_{x,i}^{\eta}\ln\beta_x\right) + \frac{1}{2}v\sum_{x=1}^{n}\sum_{i=1}^{c}\left(\mu_{x,i}^{m}\tanh\alpha_i + t_{x,i}^{\eta}\tanh\beta_x\right) \quad (21)$$

The centroid of *i*th cluster is calculated in the similar way as the definition in Eq. (18). The final objective function is presented in equation (21).

3.4. Clustering Enhancement using Repulsion

In the above described clustering technique, objective function is truly minimized only if all the centroids are identical (coincident), since the typicality of a point to a cluster, depends only on the distance between the point to that cluster.

The usage of repulsion aims to minimize the intracluster distances, while maximizing the intercluster distances, without using implicitly the restriction, but by adding a cluster repulsion term to the objective function.

$$J_{MFPCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m}w_{ji}^{m}d^{2m}(x_j v) + t_{ij}^{\eta}w_{ji}^{\eta}d^{2\eta}(x_j,v_i)) - \frac{1}{2}v\sum_{x=1}^{n}\sum_{i=1}^{c}\left(\mu_{x,i}^{m}\ln\alpha_i + t_{x,i}^{\eta}\ln\beta_x\right) + \frac{1}{2}v\sum_{x=1}^{n}\sum_{i=1}^{c}\left(\mu_{x,i}^{m}\tanh\alpha_i + t_{x,i}^{\eta}\tanh\beta_x\right) + \sum_{i=1}^{c}\eta_i\sum_{k=1}^{c}(1-u_{ik})^m + \gamma\sum_{i=1}^{c}\sum_{k=1,k\neq i}^{c}\frac{1}{d^2(v_j,v_k)} \quad (22)$$

Where γ is a weighting factor, and u_{ik} satisfies:

$$u_{ik} \in [0,1], \forall i \quad (23)$$

The repulsion term is relevant if the clusters are close enough. With growing distance it becomes smaller until it is compensated by the attraction of the clusters. On the other hand, if the clusters are sufficiently spread out and the intercluster distance decreases, the attraction of the cluster can be compensated only by the repulsion term.

Minimization of objective function with respect to cluster prototypes leads to:

$$v_i = \frac{\sum_{j=1}^{n}u_{ij}x_j - \gamma\sum_{k=1,k\neq i}^{c}v_k\frac{1}{d^2(v_k,v_i)}}{\sum_{j=1}^{n}u_{ij} - \gamma\sum_{k=1,k\neq i}^{c}\frac{1}{d^2(v_k,v_i)}} \quad (24)$$

Singularity occurs when one or more of the distances $d^2(v_k,v_i) = 0$ at any iteration. In such a case, v_i cannot be calculated. When this happens, assign zeros to each nonsingular class (all the classes except i) and assign 1 to class i, in the membership matrix U.

An alternative repulsion term for (22) in order to minimize the objective function is given by

$$J_{MFPCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij}^{m} w_{ji}^{m} d^{2m}(x_j,v) + t_{ij}^{\eta} w_{ji}^{\eta} d^{2\eta}(x_j,v_i)) - \frac{1}{2}\nu\sum_{x=1}^{n}\sum_{i=1}^{c}\left(\mu_{xi}^{m}\ln\alpha_i + t_{xi}^{\eta}\ln\beta_x\right) + \frac{1}{2}\nu\sum_{x=1}^{n}\sum_{i=1}^{c}\left(\mu_{xi}^{m}\tanh\alpha_i + t_{xi}^{\eta}\tanh\beta_x\right) + \Sigma_{i=1}^{c}\eta_i\Sigma_{k=1}^{n}(1-u_{ik})^m + \gamma\Sigma_{i=1}^{c}\Sigma_{k=1,k\neq i}^{c} e^{-d^2(v_k,v_i)} \quad (25)$$

The weighting factor γ is used to balance the attraction and repulsion forces, i.e., minimizing the intradistances inside clusters and maximizing the interdistances between clusters.

3.5. *Cluster Validity Index for the Proposed Clustering Technique*

This section provides a validity index for the proposed clustering. Let X = {x_1, . . .,x_n} be a data set in R_s. Let μ= {μ_1. . . μ_c} is a partitions based on the proposed clustering algorithm.

This paper considers two factors with a normalized partition coefficient and an exponential separation measure to validate every cluster. Next the author used these two terms to generate a new validity index, called a partition coefficient and exponential separation (PCAES) index. The PCAES index for cluster i is defined as

$$\mathrm{PCAES}_i = \sum_{j=1}^{n}\mu_{ij}^2/\mu_M - \exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\right) \quad (26)$$

Where

$$\mu_M = \min_{1\le i\le c}\left\{\sum_{j=1}^{n}\mu_{ij}^2\right\} \text{ and } \beta_T = \frac{\Sigma_{i=1}^{c}\|\alpha_1 - \bar{\alpha}\|^2}{c} \quad (27)$$

The term of a normalized partition coefficient (NPC) is used with

$$\sum_{j=1}^{n}\mu_{ij}^2/\mu_M \quad (28)$$

to compute the compactness for the cluster i comparative to the most compact cluster which has the compactness measure lM. This term is similar to the compactness measure for cluster i used in the PC index, where the measure is taken as an average, not as a relative value. The compactness value in (28) will belong to the interval (0, 1].

The exponential-type separation measure for cluster i with

$$\exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\beta_T\right) \quad (29)$$

takes advantage of exponential function that calculates the distance between cluster i and its closest cluster. Moreover, we consider it relative to β_T of the total average distance measure for all clusters. The total average distance measure of all clusters is similar to the separation measure $K_m(\mu,a)$ defined by the FS index. The exponential function is taken to make the separation measure in the interval (0, 1] and also make the compactness (28) and separation (29) to have the same range (or degree) of measurement. Another motivation for taking the exponential function is that an exponential operation is highly useful in dealing with the classical Shannon entropy and cluster analysis. Especially, an exponential-type distance gives robust property based on the influence function analysis.

Since the compactness and separation for each cluster are restricted on

$$0 < \sum_{j=1}^{n}\mu_{ij}^2/\mu_M \le 1 \quad (30)$$

and

$$0 < \exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\beta_T\right) \le 1 \quad (31)$$

Next the boundary for PCAES_i is

$$-1 < \mathrm{PCAES}_i < 1 \text{ for all } i = 1, \ldots, c \quad (32)$$

It can be observed that the proposed validity criterion PCAES_i could detect each cluster with two measures from a normalized partition coefficient and an exponential separation. The large PCAES_i value means that the cluster i is compact inside and separated from the other (c - 1) clusters. The small or negative value of PCAES_i indicates that cluster i is not a well-identified cluster. Finally, the PCAES validity index is then defined as

$$PCAES(c) = \sum_{i=1}^{c} \text{PCAES}_i = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^2/\mu_M - \sum_{i=1}^{c} exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\beta_T\right) \quad (33)$$

Obviously,

$$-c < \text{PCAES}(c) < c \quad (34)$$

In the validity index, initially PCAES_i is used to measure the compactness and separation for each cluster and then summed all PCAES_i as PCAES(c) to measure the compactness and separation for the data structure. Thus, the total compactness of the data set is measured by the term

$$\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^2/\mu_M \quad (35)$$

which is the normalized PC index and the total separation of the data set is measured by the term

$$\sum_{i=1}^{n} \exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\beta_T\right), k = 1, \ldots, c \quad (36)$$

The large PCAES(c) value means that each of these c clusters is compact and separated from other clusters. The small PCAES(c) value means that some of these c clusters are not compact or separated from other clusters. The maximum of PCAES(c), with respect to c, could be used to detect the data structure with a compact partition and well-separated clusters. Thus, an optimal c* can be found by solving $\min_{2 \le c \le n}$ PCAES(c) to produce a best clustering performance for the data set X.

The consideration of normalizing the partition coefficient can give us a small PCAES_i, value when cluster i contains only a few points and the index PCAES will be then relatively small. This gives us an alarm whether noisy points are taken into compact and separated clusters or not. This situation often occurs in real applications. Other indexes do not own this property. Thus, using the proposed validity index not only gives us an optimal cluster number estimate, but also presents more information about the data structure.

IV. Experimental Results

The proposed approach for clustering unlabeled data is experimented using the two benchmark datasets which are Iris and Wine Dataset from the UCI machine learning Repository. All algorithms are implemented under the same initial values and stopping conditions. The experiments are all performed on a GENX computer with 2.6 GHz Core (TM) 2 Duo processors using MATLAB version 7.5.

Experiment with Iris Dataset

The Iris flower data set (Fisher's Iris data set) is a multivariate data set. The dataset comprises of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from every sample; they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, Fisher has developed a linear discriminant model to distinguish the species from each other. It is used as a typical test for many classification techniques. The proposed method is tested first using this Iris dataset. This database has four continuous features consisting of 150 instances: 50 for each class.

The mean square error (MSE) of the centers $MSE = \sqrt{\|v_c - v_t\|^2}$ where v_c is the computed center and v_t is the true center. The cluster centers found by the proposed technique are closer to the true centers, than the centers found by other clustering techniques. The mean square error for the cluster centers for the different approaches are presented in table I.

TABLE I

MEAN SQUARE ERROR VALUE OBTAINED FOR THE THREE CLUSTERS IN THE IRIS DATASET

	MFPCM	**PCFPCM**	**PCFPCM with Repulsion**
Cluster 1	0.3215	0.1014	0.0785
Cluster 2	0.4127	0.2147	0.1258
Cluster 3	0.3121	0.1019	0.0758

From the experimental observations it can be found that the proposed approach produces better cluster than the existing fuzzy approaches. The MSE value is highly reduced which represents better clustering.

Next the proposed PCAES index is compared with the other seven indexes PC, PE, MPC and FS. We implemented the proposed clustering algorithm on each data set with the cluster number c = 2 to 8.

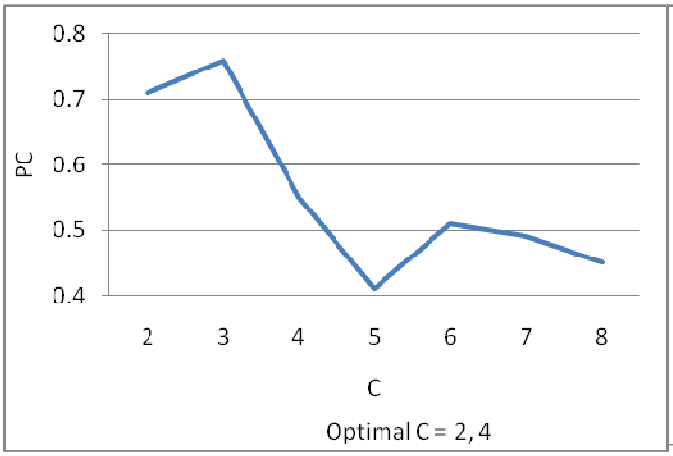

Figure 1: Validity Index using PC

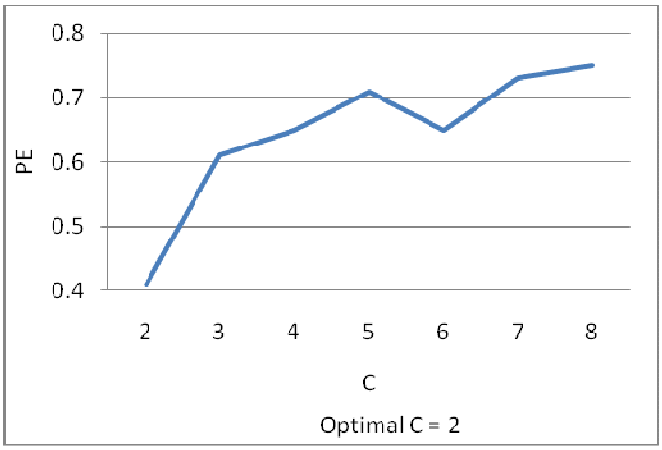

Figure 2: Validity Index using PE

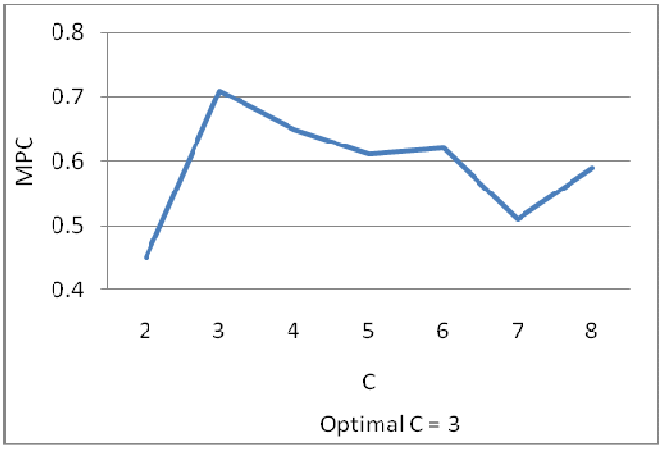

Figure 3: Validity Index using MPC

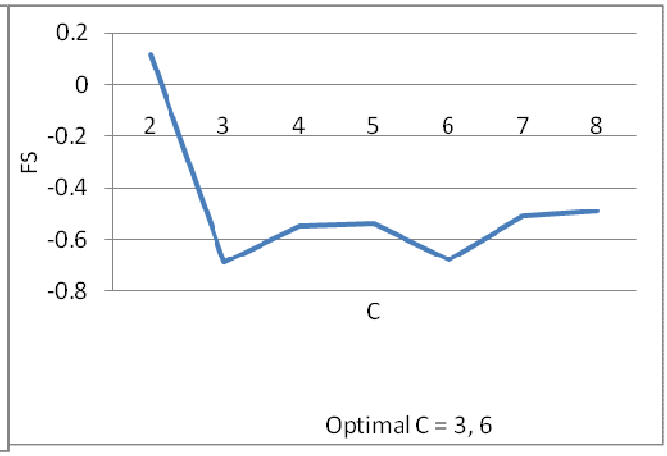

Figure 4: Validity Index using FS

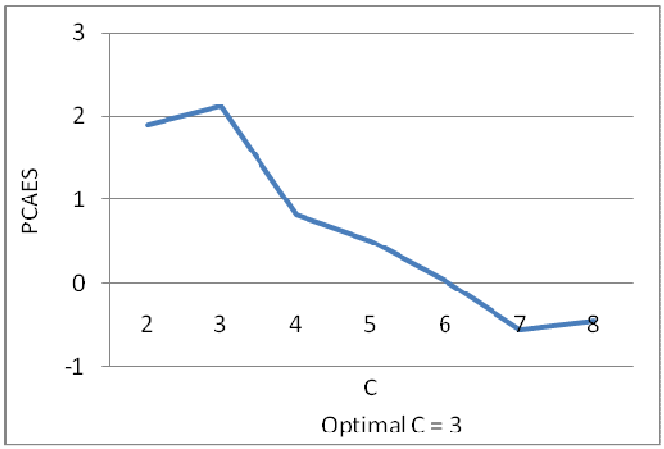

Figure 5: Validity Index using PCAES

The usage of validity index will be helpful in better partitioning of data with better accuracy. The validity index resulted for various techniques like PC, PE, MPC and FS is provided in figures 1, 2, 3 and 4 respectively. From the figures it can be clearly observed that all the existing indexing techniques cannot be able to predict the value of c effectively for achieving the better clustering result. The validity index for the PCAES index is provided in figure 5 in which the optimal value for c is found to be 3.

The validity index obtained for various techniques in case of iris dataset is provided in table 3. From the table it can be observed that the optimal c value resulted for using PC is 2 & 4, for PE is 2, for MPC is 3 and for FS is 3 & 6. Whereas by using the proposed PCAES technique, the optimal c

value resulted is 3 which will be the better value for c to result in accurate clustering.

TABLE III

VALUES OF VALIDITY INDEXES FOR VARIOUS METHODS

C	PC	PE	MPC	FS	PCAES
2	0.71	0.41	0.45	0.12	1.89
3	0.76	0.61	0.71	-0.69	2.12
4	0.55	0.65	0.65	-0.55	0.82
5	0.41	0.71	0.61	-0.54	0.51
6	0.51	0.65	0.62	-0.68	0.02
7	0.49	0.73	0.51	-0.51	-0.56
8	0.45	0.75	0.59	-0.49	-0.45

V. CONCLUSION

The problem of clustering is solved in this paper. This paper uses Fuzzy Possibilistic C-Means (FPCM) algorithm which merges the advantages of both fuzzy and possibilistic c-means technique. Then a modification is made to the objective function of FPCM to results in better performance. Then the Penalized and Compensated constraints are used in the objective function. Next, the intercluster distance is reduced by using Repulsion terms in objective function. Finally for determining the number of clusters, Partition Coefficient and Exponential Separation (PCAES) method is employed. This technique uses the factors from a normalized partition coefficient and an exponential separation measure for every cluster and then combines these two factors to create the PCAES validity index. The experimental result shows that the proposed clustering algorithm results in better accuracy when compared to the existing clustering algorithms. In work can be extended in future by modifying the objective function with the help of various constraints.

REFERENCES

[1] 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 , Issue: 4, Publication Year: 2005 , Pp: 517 – 530

[2] Vuda. Sreenivasarao and Dr. S. Vidyavathi, "Comparative Analysis of Fuzzy C- Mean and Modified Fuzzy Possibilistic C -Mean Algorithms in Data Mining", IJCST Vol. 1, Issue 1, September 2010, Pp. 104-106

[3] Xiao-Hong Wu and Jian-Jiang Zhou, "Possibilistic Fuzzy c-Means Clustering Model Using Kernel Methods", International Conference on Intelligent Agents, Web Technologies and Internet Commerce, Vol. 2, Publication Year: 2005 , Pp. 465 - 470

[4] Ojeda-Magafia, B.; Ruelas, R.; Corona-Nakamura, M.A. and Andina, D., "An Improvement to the Possibilistic Fuzzy c-Means Clustering Algorithm", Automation Congress, 2006. WAC '06, Publication Year: 2006 , Pp. 1 – 8

[5] Yang Yan and Lihui Chen, "Hyperspherical possibilistic fuzzy c-means for high-dimensional data clustering", ICICS 2009. 7th International Conference on Information, Communications and Signal Processing, 2009, Publication Year: 2009 , Pp. 1-5

[6] Long Yu; Jian Xiao and Gao Zheng, "Robust Interval Type-2 Possibilistic C-means Clustering and its Application for Fuzzy Modeling", FSKD '09. Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, Vol: 4, Publication Year: 2009 , Pp. 360 – 365

[7] Chunhui Zhang, Yiming Zhou and Trevor Martin, "Similarity Based Fuzzy and Possibilistic c-means Algorithm", Proceedings of the 11th Joint Conference on Information Sciences (2008), Pp. 1-6

[8] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981

[9] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York, Plenum, 1981.

[10] M. Barni, V. Cappellini, A. Mecocci, “Comments on A possibilistic approach to clustering", IEEE Trans on Fuzzy Systems, vol. 4: 393-396, 1996.
[11] M R. Berthold and D. J. Hand, “Intelligent Data Analysis”.Springer-Verlag, Berlin, Germany 1999.
[12] M. W. Berry, “Survey of Text Mining”, Springer-Verlag, New York, NY, USA 2003.
[13] N. R. Pal, K. Pal and J. C. Bezdek, “A mixed c-means clustering model”, Proceedings of the Sixth IEEE International Conference on Fuzzy Systems, Vol. 1, pp. 11-21, Jul. 1997.
[14] K. Lung, “A cluster validity index for fuzzy clustering”, Pattern Recognition Letters 25(2005) 1275-1291.
[15] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, “Advances in Knowledge Discovery and Data Mining”, AAAI Press and MIT Press, Menlo Park and Cambridge, MA, USA 1996.
[16] M.S. Yang, On a class of fuzzy classi4cation maximum likelihood procedures, Fuzzy Sets and Systems 57 (1993) 365–375.
[17] M.S. Yang, C.F. Su, On parameter estimation for normal mixtures based on fuzzy clustering algorithms, Fuzzy Sets and Systems 68 (1994) 13–28.
[18] J.S. Lin, Fuzzy clustering using a compensated fuzzy hopfield network, Neural Process. Lett. 10 (1999) 35–48.
[19] W.L. Hung, M Yang, D. Chen,” Parameter selection for suppressed fuzzy c-means with an application to MRI segmentation”, Pattern Recognition Letters 2005.
[20] Juan Wachs, Oren Shapira and Helman Stern, "A Method to Enhance the Possibilistic C-Means with Repulsion Algorithm based on Cluster Validity Index", Advances in Intelligent and Soft Computing, Springerlink, Vol. 34, Pp. 77-87, 2006.
[21] Kuo-Lung Wu and Miin-Shen Yang b, "A cluster validity index for fuzzy clustering", Elsevier Pattern Recognition Letters, Pp. 1275–1291, 2005.

D. Vanisri has received the Master of Science in Mathematics in 2001 from Madurai Kamaraj University. Then she completed her Master of Philosophy in Mathematics in the year 2003. Now she is doing research in the field of Fuzzy clustering and rule mining at Mother Terasa Women’s University, Kodaikannal. Currently she is working as a Lecturer in the Department of Computer Technology and Applications, Kongu Engineering College, Tamilnadu. She has presented many papers in national and international conferences and also guided many UG projects. She has published 3 papers in international journals.

Dr. C.Loganathan qualified basically with B.Sc and M.Sc in Mathematics in 1978 and 1980 respectively from Madras University and subsequently with M.Phil and Ph.D in Mathematics from Bharathiar University has served in various capacities as faculty member and Head of the Department of Mathematics at Kongu Engineering College, Perundurai for more than a decade. He is at present working as Principal, Maharaja Arts and Science College, Coimbatore. His unquenchable thirst for academic achievements had culminated in the publication of series of research papers, numbering more than 12 in the leading-referred national and international journals. As a research guide, he has produced many M.Phil and Ph.D candidates. He is a reviewer of many referred international journals. His areas of interest encompass Applied Mathematics, Control Theory, Numerical Methods, Quantitative Techniques and Neural Networks. He has co-authored the books on “Quantitative Methods in Management, Engineering Mathematics I and Engineering Mathematics II”.