International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 3, Issue 5, September-October 2014 **ISSN 2278-6856**

Improved Membership Function for Multiclass Clustering with Fuzzy Rule Based Clustering Approach

Archana N. Mahajan[1], Prof. Dr. Girish Kumar Patanaik[2] , Sandip S. Patil[3]

[1]Research Scholar, S.S.B.T.'s College of Engg.and Technology,
Bambhori, Jalgaon, Maharashtra

[2]Professor and H.O.D. Computer Engg. Department, S.S.B.T.'s College of Engg. and Technology,
Bambhori, Jalgaon, Maharashtra

[3]Associate Professor Computer Engg. Department, S.S.B.T.'s College of Engg. and Technology,
Bambhori, Jalgaon, Maharashtra

Abstract

Fuzzy clustering is the combination of clustering and fuzzy set theory. It is useful to handle the problem of determining the vague boundaries of clusters. Fuzzy clustering is better than Crisp clustering when the boundaries between the clusters are vague and ambiguous. In both fuzzy and crisp clustering algorithms there is need and requirement to know the number of potential clusters and/or their initial positions in advance. The existing system identifies the potential clusters in given dataset by itself. It uses the fuzzy rules for identifying the potential clusters. When the multiclass dataset is given as input to the existing system, the number of clusters discovered in the multiclass dataset are less than the classes in it because it forms the two class problem of the given data and applies it as input to genetic algorithm for fuzzy rule generation. The genetic algorithm generates the more general rules. Since each class represent some aspects of that particular class and it can be used to generate the taxonomy, it becomes essential to find optimized clusters. The proposed system identifies the number of clusters which are equal to the number of implicit classes in multi class data. There are three main phases in the proposed system. First, it preprocesses the multi class data. In the second phase, it generates the fuzzy rules with subtractive clustering. The best membership function for multi class data is searched in third phase with the minimum Euclidean norm. The proposed system, then finds the number of clusters in the given dataset with the best membership function and the fuzzy rules generated. The number of clusters is also optimized using the adaptive network based fuzzy inference system. These clusters are discovered without prior knowledge about the number of classes available in the given data. The proposed system can be used to generate the class labels and identify the sub-homogeneous patterns in the given data.

Keywords: Fuzzy Clustering, membership function, Fuzzy Rules, Fuzzy Rule Based Clustering

1. INTRODUCTION

Data Mining is defined as a process of distinguishing novel, useful and understandable patterns in data. Novel means the pattern is not known earlier. Useful means inference can be devised from the patterns. Understandable means one can interpret and figure out the pattern. The need of data mining arises due to the growth of data, data warehouse and computational power of the computers. Volume and dimensionality of the data has increased, due to this the human analysis skills for data analysis are becoming inadequate. So there is need of data mining techniques. Data mining [1] is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The process of identifying patterns in data mining can be clustering or classification. Classification is called as a supervised learning method as the class labels are available. Clustering\index is called as unsupervised learning because class labels are not known in advance. The goal of clustering is descriptive. The goal of classification is predictive. Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is essential. In classification tasks, an important part of the assessment is extrinsic, because the groups must reflect some reference set of classes. The rest of this paper is organized as follows: Section 2 describes the fuzzy logic and clustering. The literature survey and literature summary of fuzzy clustering is presented in Section 3. Section 4 describes the implementation of the proposed solution. Result and discussion are presented in Section 5. Section 6 presents the conclusion of the paper.

2. FUZZY LOGIC AND CLUSTERING

In many real world application areas, knowledge is represented in terms of imprecise linguistic words from a natural language. A linguistic variable means a variable whose values are words or sentences in a natural or artificial language. For example, honesty is a linguistic variable. The linguistic values of this variable can be not honest, sometime honest, very honest and extremely honest. Fuzzy logic is introduced by Zadeh [2] as a way of representing and manipulating data that is not exact, but

rather imprecise. Fuzzy logic [2] is the technique to represent approximate reasoning. Approximate reasoning is a mode of reasoning which is not exact or not very inexact. Fuzzy logic gives a realistic framework for human reasoning. It is a superset of Boolean logic that has been extended to handle, approximate reasoning or the partial truth values. Partial truth values are values between "completely true" and "completely false". Probability can also be treated as linguistic variables with values such as likely, very likely, unlikely. The probability based linguistic variables can also be presented through fuzzy logic. Fuzzy clustering [2] is the mixture of fuzzy logic and clustering, which is the requirement of modern computing. The aim of fuzzy clustering is to model the ambiguity within the unlabeled data objects efficiently. Every data object is assigned a membership to represent the degree of belonging to a certain class. The requirement that each object is assigned to only one cluster is relaxed to a weaker requirement in which the object can belong to all of the clusters with a certain degree of membership. Thus, it assigns degrees of membership in several clusters to each input pattern(data object). A fuzzy clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership. Fuzzy clustering is able to find out the relation between each object and the revealed clusters. It becomes simple to construct the rule based fuzzy model which is more understandable to human, due to fuzzy clustering. Fuzzy clustering can be categorized in three categories: Fuzzy relation based clustering, objective function based clustering and fuzzy rule based clustering.

3. LITRATURE SURVEY

Fuzzy clustering is used to model the ambiguity. It can be categorized into three different types: Fuzzy relation based clustering, objective function based clustering and fuzzy rule based clustering. 1. Fuzzy Clustering based on Fuzzy Relation [3]: This clustering method is based on fuzzy relations. It includes an N-step procedure by using the composition of fuzzy relations beginning with a reflexive and symmetric fuzzy relation R with X. The data set is partitioned into the number of clusters by equivalence relation R λ. 2. Fuzzy Clustering based on Objective Function [4]: The objective function is used as the dissimilarity measure for fuzzy clustering in the objective function based fuzzy clustering methods. 3. Fuzzy Rule based Clustering [4]: The fuzzy rules are generated for the given input data set and by generating fuzzy rules the inference is drawn to form the clusters. It considers data objects in the fuzzy subspace of the best rule as the first clusters members and discards them from the problem space. The Fuzzy Rule Based clustering(FRBC)repeats this process to extract all possible clusters. By the assignment of some distinct labels to the obtained fuzzy rules, which represent the clusters, and classification of the original data from them, the actual boundaries of the clusters can be identified. The members of these clusters can then be used to group the uncovered data objects. Liang et al., in [5], introduced Cluster analysis based on fuzzy equivalence relation. The approach used is the distance measure between two trapezoidal fuzzy numbers is used to aggregate subject's linguistic assessments. The distance measure is used to characterize the inter-objects similarity. The linguistic assessment is for attributes ratings to obtain the compatibility relation. Then a fuzzy equivalence relation based on the fuzzy compatibility relation is constructed. The algorithm then uses a cluster validity index L to determine the best number of clusters. Then a suitable λ-cut value is taken based on the fuzzy compatibility relation. The cluster analysis is done through the λ-cut value and a cluster validity index. This algorithm finds out the best number of clusters and the objects belonging to the clusters. The cluster validity index measure is defined as:

$$L = T/n * (d\min)2$$

Where L is a cluster validity index, n is the number of objects, T is a normalized compatibility relation. Miin-Shen Yang and Hsing-Mei Shih, in [6], introduced cluster analysis based on fuzzy relations. The algorithm creates the max-t compositions from max-min n-step procedure. Amax-t similarity-relation matrix is obtained by beginning with a proximity-relation matrix based on the max-t n-step procedure. Then a clustering algorithm is used for any max-t similarity-relation matrix. This algorithm is also able to process the missing data. The max-t similarity relation is defined as

$$R = [r]_{n\times n}$$

Where R is max-t similarity relation, r is max-t transitivity.

Objective function based clustering: The objective function is used to form the clusters in fuzzy clustering. Wang et al., in [7], introduced improving fuzzy, c-means clustering based on feature-weight learning algorithm. The approach used by the algorithm is the featureweight as the dissimilarity measure for the objective function. Each feature within the input space is assigned a weight. The algorithm improves the performance of Fuzzy C-means clustering by integrating the feature weight matrix with the Euclidean distance known as a weighted similarity measure. Dao-Qiang Zhang and Song-Can Chen, in [8], introduced clustering incomplete data using kernel-based fuzzy c-means algorithm. The approach used in the algorithm is the kernel function is created. Kernel function is derived from the Gaussian objective function and the hyper tangent function. The kernel function minimizes the basic objective function in fuzzy c-means. The kernel based objective function increases the capability of fuzzy c-means to cluster the incomplete data. Nikhil R. Paul and James C. Bezdek, in [9], introduced on cluster validity for the Fuzzy C-means model. The approach used in the algorithm is analysis done on the weighting exponent m of the fuzzy C means clustering, which is the core factor for the fuzzy partitions. The Low and high value of weighting exponent influences the validity indexes of the membership function used. Fuzzy rule based clustering:

The Fuzzy rules are the basis for the formation of the clusters.Jun et al., in [10], introduced a Method for Fish Diseases Diagnosis Based on Rough Set and FCM Clustering Algorithm. The approach used in the algorithm is combination of rough set and Fuzzy C-means (FCM) clustering. The rough set is used for acquiring knowledge and for the formation of the decision-making table. The Redundant properties and samples are removed and then the Fuzzy C means clustering is applied. Mehrnoosh Sinaee and Eghbal G. Mansoori, in [11], introduced fuzzy Rule Based Clustering for Gene Expression Data. The approach used in the algorithm is grouping of the similar genetic data is done based on the fuzzy rules generated. Then the Gene Expression data is again applied as the input and the clusters are formed. The fuzzy rules generated are understandable to the human user. Eghbal G. Mansoori, in [12], introduced a fuzzy rule based clustering Algorithm. The approach used in the algorithm is the generation of fuzzy rules with the help of genetic algorithm and then using these rules the clusters are formed. Each rule is used as the cluster formation measure. The algorithm generates the fuzzy rules which are human understandable. It applies a supervised classification approach to do the unsupervised cluster analysis. It tries to automatically explore the potential clusters in the data patterns with no prior knowledge. Setnes et al., in [13], introduced Rule-Based Modeling. The approach used in the algorithm is first the TS (Takagi Sugeno) rule based model is constructed. The TS model construction is done in two phases: 1) Structure Identification and 2) Parameter estimation. In the structure identification step, the antecedent and consequent variables of the model are determined. From the available data sequences, a regression matrix X and an output vector y are constructed. In the parameter estimation step, the number of rules K, the antecedent fuzzy sets Aij, and the parameters of the rule consequents ai; bi for i = 1, 2,.....,K are determined. The rules are generated by product-space clustering. The algorithm is useful for achieving precision along with a good qualitative description in linguistic terms. Delgado et al., in [14], introduced a Fuzzy Clustering-Based Rapid Prototyping for Fuzzy Rule-Based Modeling. The approach used in the algorithm is fuzzy clustering process for building the rapid prototype, for the generation of the approximation of a fuzzy model. The objective of a rapid methodology is to obtain model to use as a first approximation of fuzzy model for systems. Stephen L. Chiu, in [15], introduced fuzzy model identification based on cluster estimation. The approach in the algorithm is a cluster center estimation method is used for fuzzy rule formation. This algorithm works well by providing the similar degree of accuracy and robustness of noisy data. The algorithm also reduces computational complexity.
Table I presents the literature summary.

TABLE II: Literature Summary

Author	Technique	Datasets	Performance Metric
Liang et al. [5]	Fuzzy Equivalence Relation	Strategic Data (Air-freight forwarder data withSAS)	uses cluster validity indexmeasure $L = \frac{T}{n} * (dmin)2$
Miin Shen Yang and Hsing-Mei Shih [6]	Fuzzy relations extended Tamura's max-min n stepprocedure	Portraits data	max-t similarity relation $R = [\gamma]_{n\times n}$
Wang et al. [7]	feature-weight learning	Bupa Bostan Iris	Weighted similarity measure $\rho_{ij}^{w} = \frac{1}{1+\beta * d_{ij}^{w}}$
Dao-Qiang Zhang and Song-Can Chen [8]	kernel function based FCM	Iris	kernel-induced metric $d(x,y) \triangleq \| \Phi(x) - \Phi(y) \|$
Nikhil Paul and Bezdek [9]	weighting exponent m based cluster validity	Iris	weighting exponent m core-deciding factor $C_{FCM} = [X; (c, m, P_{i3}, \ldots P_{i7})]$
jun and et al. [10]	Rough set as decision Table	fish disease diagnosis record	Subject Measurement Matrix
M. Sinaee and Mansoori [11]	Fuzzy Rule generation using GA	All-Aml Leukemia, Central Nervous System	compatibility grade $C_j = \frac{\sum_{x_p \in G_j} \mu_j(x_p) \times x_p}{\sum_{x_p \in G_j} \mu_j(x_p)}$
Eghbal G. Mansoori [12]	Fuzzy Rule based clus-Tering	Iris Ecoli Wine	compatibility grade $C_j = \frac{\sum_{x_p \in G_j} \mu_j(x_p) \times x_p}{\sum_{x_p \in G_j} \mu_j(x_p)}$
Setnes et al. [13]	Fuzzy Rule base con-struction through mea-surements	N/A	regression matrix X
Delgado et al. [14]	Rapid prototype formation for approximati-on	N/A	Estimation procedure formula $\hat{y} = \frac{\sum_{h=1}^{k} \mu_{ch}(\tilde{x}) * f_h(x)}{\sum_{h=1}^{k} \mu_{ch}(x)}$
Stephen L. Chiu [15]	estimation of cluster center for fuzzy rule generation	N/A	Potential of each data pointis given as $P_i \Rightarrow P_i - P_k^* e^{-\beta \|x_i - x_k^*\|^2}$

4. IMPLEMENTATION

The proposed solution first pre-processes the multi class data. The pre-processed data is used to generate the fuzzy rules. Then the best membership function is searched. The basic set of fuzzy rules and the best membership function is given as input to the adaptive network based fuzzy clustering for the cluster formation. The assumption used for the proposed system is that the input has to be multi-class data of numeric values only.

a. Fuzzy Rules

Fuzzy rules are a collection of linguistic statements that describe how the fuzzy inference system should make a decision regarding classifying an input or controlling an output. There are two parts in the fuzzy if then rules. The 'if' part is called as the antecedent part and the 'then' part is known as the consequent part. The antecedent part draws inference and gives decision about the consequent part. Fuzzy IF–THEN Rule [12] The fuzzy IF-THEN rule for a clustering problem with n attributes can be written as follows:

Rule R_j: IF x_1 is A_{j1} and . . . and x_n is A_{jn} THEN class

$C_j \forall j$

Where

j = 1, . . . , N

$X = [x_1, x_2, . ., x_n]$ is an n-dimensional pattern vector

A_{ji} (i = 1,…, n) is an antecedent linguistic value of R_j

C_j is the consequent class

N is the number of fuzzy rules

Generally, for an M-class problem with m labelled data objects $X_p = [x_{p1}; x_{p2},…, x_{pn}]$, p = 1, . . . ,m, the task to design the classifier is to generate a set of N fuzzy rules.

The examples of fuzzy rules are:

If x_1 is A_1 and x_2 is B_1 then class C_1
If x_1 is A_2 and x_2 is B_1 then class C_2
If x_1 is A_1 and x_2 is B_2 then class C_3
If x_1 is A_2 and x_2 is B_2 then class C_4

b. Membership Function

In fuzzy logic, each natural-language word is described by a membership function $\mu(x)$. A membership function [16] that assigns, to every number x, the degree μ (x) $\varepsilon[0, 1]$ to what this number satisfies the corresponding property (e.g. the degree to which the number x is small). A membership function is used to define the imprecise (fuzzy) properties. A membership function provides a measure of the degree of similarity of an element to a fuzzy set. Each membership function [16] states skilled opinion to get proper numerical values for fuzzy properties. The fuzzy clustering uses the fuzzy membership function to find the relationship between objects. There are various membership functions available such as absolute membership, probabilistic membership, Gaussian mixture, etc. In fuzzy Rule Based Clustering, the pattern space is partitioned into fuzzy subspaces, and each subspace is identified by a fuzzy rule, if there are some patterns in that subspace. To do partitioning, generally, K suitable membership functions are used to assign K linguistic values to each input attribute. Generally, triangular membership functions are used for this purpose, because they are simpler and more human understandable. With the help of the membership function, we can assign the linguistic labels to the patterns such as small, medium, large. Membership function is the way of characterizing the fuzziness. The performance of a fuzzy system depends on the rule base and its membership functions. With a given rule base, the membership functions can be optimized in order to improve the overall performance of a fuzzy system.

4.3 Basic Design Architecture

Figure 1 shows the basic design architecture of the proposed system. It includes the seven main blocks and the proposed system's input and the output. The arrows in Figure show the flow within the system.

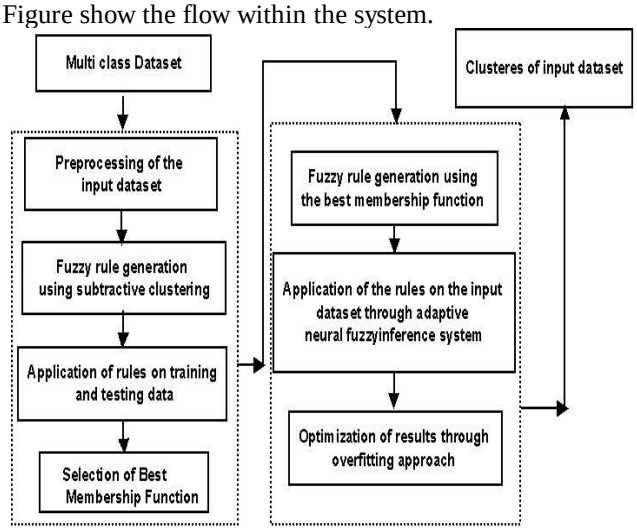

Figure 1 Basic Design Architecture

The input to the proposed system is multi-class dataset and the output is the number of clusters in the multi class dataset. Since each fuzzy inference system requires input data in the particular format so the original dataset is required to be preprocessed. The preprocessing of the data is performed by the first block. There are number of data preprocessing techniques [1] such as data cleaning, data normalization, data integration, data reduction. Preprocessing of the input dataset includes the analysis of the raw data in the input dataset. The block accepts the multi class dataset as an input. Since the dataset contains the data which is in the form of cell arrays of matrices. This cell array of matrices is converted into the single matrix. Then the size of the input data set is calculated. The number of attributes within the input dataset is calculated. So that the size of single row with the data set can be known. The data normalization is done so that all the attributes can get the equal weight. The normalization causes the scaling of the data objects. Then the training and testing data elements are selected. Thus the output of this block is the set of training and testing data elements.

Let input dataset of size m×n patterns where m is number of patterns in n-dimensional space(type of data).

$$I = \{x_{11}, x_{12}, .., x_{mn}\}$$

Where For x_{ij} ; i = 1, 2,…,m; j = 1,…, n.

Let the size of input dataset is S_z.

$$s_z = |I|$$

The normalized multidimensional single matrix [1 × n] calculated as

$$[1 \times n] = a[n]$$
$$\text{Where } a[n] = [m_1, m_{2,\dots,} m_n]_{1\times n}$$
$$m_k = \xi(i.f)$$
$$i \rightarrow [0,1,2,\dots,j]$$
$$f \rightarrow [0,1,2,\dots,j]$$
$$|f| \rightarrow 4$$

The input and output training data element of the given data set is $[m \times t]$.

$$[m \times t] = [p \times n]$$

Where p = number of rows (samples)
n = number of columns (dimensions)
Let the testing(checking) input and output data elements of the given dataset is $[a \times c] = [m \times n]$.
After the pre-processing, the second block generates the fuzzy rules using the subtractive clustering [15][17][18][19] approach. These fuzzy rules are applied on the training and the testing data with the default membership function. The next block in the proposed system is for finding the best membership functions for the input dataset so that the optimal number of cluster can be revealed by the proposed system. The various membership functions used in the proposed system are bell shaped membership function, Gaussian Curve membership function, sigmoid membership function, membership function composed of Difference between two sigmoidal membership functions, Membership function composed of Product of Two sigmoidally shaped membership functions, π-shaped membership functions, S Shaped Membership function, and Z- shaped membership function. The selected eight membership functions are applied and among them the best membership function is selected based on their Euclidean length measure. Initially all the membership functions are stored in the file. The first membership function is selected. The selected membership function is used to generate the fuzzy rules using the subtractive clustering. These fuzzy rules are applied on the input dataset through adaptive neural fuzzy inference system. Then the Euclidean length measure is calculated which drives the conclusion of goodness of fit measure. The membership function with minimum Euclidean length is selected as the best membership for the input dataset.

$$testRMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(Z_0 - Z_p)^2}$$

Where $Z_0$is the observed value
Z_p is predicted value
n is number of samples in the training data
The best membership function is defined as

$$BMF = \min(testRMSE)$$
$$FIS = f(BMF, \varsigma, [m \times t])$$

Where ς is the rule application function
$[m \times t]$ is the testing data BMF is Best Membership Function FIS is Fuzzy Inference System The last block then generates the fuzzy rules with the best membership function. These rules are applied on the input data using the adaptive fuzzy inference system [20][21][22].First the input to the block is the training and testing data set as well as the basic fuzzy inference system generated using the best membership function. The number of iterations, training error goal and initial step size are set. In the second step the algorithm uses the combination of the least square method and back propagation gradient descent method for training fuzzy inference system membership parameters to match the input training data. Euclidean length (root means square error)is calculated for the training data in the third step. The training of the networks is given for the specified number epochs (iterations). Then the testing (checking) data set is applied and the Euclidean length (root means square error)is calculated for the testing data. When the training error is decreased and testing error is increased, the overfitting technique is used to generate the optimize results. In overfitting the algorithm is iterated for specified number of iteration. The clustering capability of the proposed system is optimized by the over-fitting concept of the neural network.

4.4 Algorithms of the Proposed System

Input: A dataset with data patterns of a clustering problem.
Outputs: The number of clusters identified by the algorithm and the Euclidean length of the identified clusters.

Algorithm 1 Preprocessing of Input data set

Require: A multiclass dataset

1: Start
2: raw=xlsread (dataset.xlsx)
3: Sz=Size (raw)
4: for i=1 to Sz (1) do
5: temp = raw (i, 1)
6: temp = cell2mat (temp)
7: for j=1 to s (2) do
8: alldatin (i) = str2num (temp (1, p: p+2))
9: p= p+4;
10: end for
11: end for
12: Goto step 4 and repeat the process for next row.
13: Repeat the steps 4 to 6 for the whole dataset.
14: Select the input and output training data element randomly for the given dataset $[m \times t]$
datin = alldatin(1:2:150, :); datout = alldatout (1:2:150)
15:Select the testing (checking) input and output data elements. Chkdatin = alldatin (1:150, :); chkdatout = alldatout (1:150)
16: End Since the input data may be in any form that is it may contain noise, incomplete data objects so first the preprocessing of the input is done. First the input data set is loaded and then its' size is calculated in terms of the number of patterns in it. Then each row is selected. The number of attributes in each row is identified. The normalization of each data object is done so that each object is scaled to the specific range. The normalization process is repeated for the whole data set. Then the input

and output training data is selected by choosing random patterns in the dataset. The input and output testing data is also selected which contains the whole dataset.

Algorithm 2 Selection of Best Membership function and cluster exploration

Require: A set of eight Membership functions.

1: Start
2: Store all the membership functions in the array.
membershipfunctionsall = {gbellmf; gaussmf; sigmf; dsigmf; psigmf; pimf; smf; zmf}
3: for MF = 1:8 do
4: do
5:f1=genfis2 (datin, datout, 0.5); f2=anfis ([datin datout], fismat, [20 0 0.1]);
6: Compute the Euclidean length
fuzout2=evalfis (datin, fismat2);
trnRMSE2 (tt) =norm (fuzout2-datout)/sqrt (length (fuzout2))
chkfuzout2=devilfish (chkdatin, fismat2)
chkRMSE2 (tt) =norm (chkfuzout2-chkdatout)/sqrt (length (chkfuzout2))
7: end for
8: BMF=min(chkRMSE)
9: genfis2(datin,datout,0.5)
10: f2=anfis ([datin datout], fismat, [20 0 0.1])
11: chkfuzout2=evalfis(chkdatin, fismat2)
12: Calculate the number of clusters in the input dataset
ans = round (chkfuzout2)
ans = ans-chkdatout
ans = find (ans)
ans = size(ans)

13: End The Algorithm 2 describes the steps for selection of best membership function and cluster exploration. First all the membership functions are stored in the array. In the second step the fuzzy rules are generated using the first membership function and training data. These fuzzy rules are applied as basic fuzzy inference set to adaptive neural network based fuzzy inference system for cluster formation. The Euclidean length of training and testing data using selected membership function is calculated in the next step. This process is repeated for all the membership function. The membership function with minimum Euclidean length is selected as the best membership function. The fuzzy rules are generated using the best membership function. These rules are applied as the basic inference set to adaptive neural network based fuzzy inference system. Then euclidean length of the testing data is calculated. To determine the number of clusters, the rule application result on the testing data is rounded since it gives the number of clusters in the multiclass input dataset.

5. RESULTS AND DISCUSSION

The proposed system is evaluated via experiments on the real world data. The real world data is chosen from Irvine Machine Learning Repository. The real world data is provided as an input to the proposed system without any change in it. It is applied as it is available to the proposed system. The data is pre-processed initially. The proposed system's algorithms are implemented using the MATLAB 7.10.0(R2010a) on the Windows XP operating system. The experiments are performed on 2.00 GHz Intel Core to Duo T6400 processor with 2.99GB of RAM. To evaluate the ability of the Proposed system to explore the natural clusters in the real world data 10 classification data sets with the numerical attributes are chosen from the University of California at Irvine Machine Learning Repository [34]. These datasets with their attributes, the in-built classes and the number of samples in them are summarized in Table 2. The various data sets are Iris, Thyroid,Ecoli,Glass,vowel, wine, vehicle, WDBC, Ionosphere and Sonar.

TABLE IIII: Datasets Used for Proposed System Evaluation

Dataset	Number of Attributes	Number of Classes	Number of Samples(m)
Iris	4	3	150
Thyroid	5	3	215
Ecoli	7	8	336
Glass	9	6	214
Vowel	10	11	990
Wine	13	3	178
Vehicle	18	4	94
WDBC	30	2	569
Ionosphere	33	2	351
Sonar	60	2	208

Table II contains six multiclass datasets. The multiclass data sets are those datasets which contains more than two classes. Their clustering results by the proposed system is of more importance as the objective of the proposed system is to discover the clusters in the data pattern and identify them with complete fuzzy rules. Various multiclass datasets are Iris, Thyroid, Ecoli, Glass, Vehicle, and Vowel.

a. Experimental Results

The proposed system's performance is tested on ten real world data set as shown in Table II. The performance of proposed system is measured using the euclidean distance. The performance of the proposed system is evaluated by the training and testing data using the euclidean distance. The euclidean distance is root means square error (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(Z_0 - Z_p)^2}$$

Where Z_o is observed value, Z_p is predicted value and n is number of samples. Table III illustrates the clustering results of the proposed system that are obtained for the multiclass datasets. For example Iris, Thyroid, Ecoli, Glass, vowel, vehicle are multiclass datasets. As shown in Table III the number of clusters discovered for the multiclass datasets are equal or greater than their classes.

TABLE IVII: Results of Proposed System on Multiclass Datasets

Dataset	Actual Class Labels	Number of Cluster explored by proposed system	Euclidean length by Proposed System
Iris	3	3	0.14475
Thyroid	3	4	0.4773
Ecoli	8	12	1.1917
Glass	6	9	1.1045
Vowel	11	13	0.19866
Vehicle	4	3	1.9295

The proposed system finds more clusters than the classes in Ecoli, Thyroid, Glass, Vowel datasets. The clusters are equal to the implicit classes for Iris dataset. The proposed system's result for Vehicle data set is fewer clusters than the implicit classes.

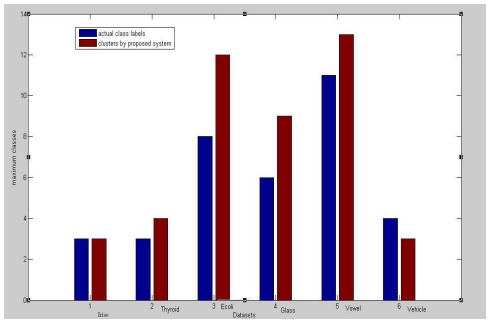

Figure 2 Clustering Results of the Proposed System on Multiclass Datasets and Actual Class Labels

Figure 2 shows the comparative analysis of the clustering results of multiclass dataset by the proposed system and actual classes in the input datasets. The clusters revealed by the proposed system are equal to the implicit classes for iris data set. The number of clusters discovered is more than their classes for Thyroid, Ecoli, Glass, and vowel. The number of cluster found is less for vehicle dataset. The clusters formed by the proposed algorithm are assembled in nature. The data analysis is the integral part of data mining. The performance of the proposed system is evaluated through the real world data set. The objective of the proposed system was to find the number of clusters in the multiclass dataset and check them whether the number of clusters are equal to the number of class labels in the dataset. The proposed system discovered the equal or more number of clusters for each multiclass dataset. This is possible because proposed system uses the hybrid approach for rule generation. Specific rules are generated by it. The specific rules represent more classes in the multi class data. This causes to have the correct number of clusters. The number of clusters is optimized by using the overfitting technique which causes correct cluster formation.

6. CONCLUSION

The performance of the clustering algorithm heavily depends on the kind of data applied to it. When the data contains number of sub-grouping then clustering algorithm must consider the influence of each and every attribute of the input data affecting the sub grouping. It is observed that the membership function varies for each and every input. If the proper membership function is selected then it forms the correct clusters and correct generation of fuzzy rules. The proposed system uses the hybrid approach. The hybrid approach contains the selection of best membership function and rule generation. The use of proposed system is to reveal the clusters when there is no prior knowledge regarding the number of clusters. It provides a way to form the knowledge base by analysing patterns and extracting the knowledge. It is effective in generation of the class labels for input data and creates the knowledge base for the classifier and the classification tool. The proposed system is also useful for taxonomy formation which organizes the observation in the input space and groups the similar objects together. The problem of finding the number of clusters in multi class data was the main concern for the proposed system. So the other cluster formation issues such as arbitrary shaped clusters formation are not considered. It requires high computational time as each membership function is selected and fuzzy rules are generated and the euclidean length is calculated for each membership function. The proposed system works well on multi class data with numerical attributes. There is scope for future work to increase the ability of the proposed system for construction of gesture recognition system and expert system based on data stream analysis. Data stream is of infinite length and evolving nature.

References

[1] Han and Kamber, "Data mining: Concepts and Techniques,"Morgan Kaufmann Publishers, pp. 47-84, 2006.

[2] L. A. Zadeh, "Fuzzy sets and logic," Fuzzy Sets, Information and Control, pp. 338-353,August 1965.

[3] E. Ruspini, "A new approach to clustering," Inform. and Control 15, pp. 22-32, 1969.

[4] M. S. Yang, "A survey of fuzzy clustering," Mathl. Comput. Modelling, vol. 18, no. 11, pp. 1-16, October 1993.

[5] G.-S. Liang, T.-Y. Chou and T.-C. Han, "Cluster analysis based on fuzzy equivalence relation," Elsevier European Journal of Operational Research, p. 160-171, 2005.

[6] M.S.Yang and H.-M. Shih, "Cluster analysis based on fuzzy relations," Elsevier FuzzySets and Systems, pp. 197-212, 2001.

[7] X.Wang, Y.Wang, and L.Wang, "Improving fuzzy c-means clustering based on feature-weight learning," Elsevier Pattern Recognition Letters 25 (2004), vol. 25, p. 1123-1132,2004.

[8] D.-Q. Zhang and S.-C. Chen, "Clustering incomplete data using kernel-based fuzzy c-means algorithm," Springer Neural Processing Letters, vol. 18, no. 3, pp. 155-162,2003.

[9] N. R. Pal and J. C. Bezdek, "On cluster validity for fuzzy c-means clustering," IEEEtransaction on fuzzy system, vol. 3, no. 3, pp. 370-380, August 1995.

[10] X. Miao-jun, Z. Jian-ke, and L. Hui, "A method for fish diseases diagnosis based on rough set and fcm clustering algorithm," IEEE 2013 Third International Conference onIntelligent System Design and Engineering Applications, pp. 1-5, 2013.

[11] M. Sinaee and E. G. Mansoori, "Fuzzy rule based clustering for gene expression data," IEEE Computer Society 2013 4th International Conference on Intelligent Systems, Modelling and Simulation, pp. 1-5, August 1995.

[12] E. G. Mansoori, "FRBC:A Fuzzy Rule-Based Clustering Algorithm," IEEE Transactions OnFuzzy Systems, vol. 19, no. 5, pp. 960-971, October 2011.

[13] M. Setnes, R. Babuska, and H. B. Verbruggen, "Rule-based modeling: Precision and transparency," IEEE Transactions on Systems, Man, And Cybernetics Part C: Applications and Reviews, vol. 28, no. 1, pp. 165-170, February 1998.

[14] M. Delgado, A. F. Gomez-Skarmeta, and F. Martin, "A fuzzy clustering-based rapid prototyping for fuzzy rule-based modeling,"IEEE Transactions on Fuzzy Systems, vol. 5, no. 2, pp. 223-234, MAY 1997.

[15] S. L. Chiu, "Fuzzy model identification based on cluster estimation,"Journal of intelligent and fuzzy systems, vol. 2, pp. 267-278, 1994.

[16] A. Barua, L. Mudunuri and O. Kosheleva, "Why trapezoidal and triangular membership functions work so well: Towards a theoretical explanation," Journal of UncertainSystems, vol. 8, no. X, pp. 1-5, October 2014.

[17] A. Priyono, M. Ridwan, A. J. Alias, and R. Atiq, "Generation of fuzzy rules with subtractive clustering," University Teknologi Malaysia, p. 143-153, 2005.

[18] Archana Mahajan, Dr. Girish Kumar Patnaik and Sandip S. Patil, " Improvement in membership function for multiclass clustering with Fuzzy Rule Based Clustering Approach" , International Journal of Advanced Engineering and Global Technology Vol-2, Issue-4, April 2014.

[19] Atul S. Chaudhari, Dr. Girish K. Patnaik and Sandip S. Patil, "Implementation of Minutiae Based Fingerprint Identification System using Crossing Number Concept" , International Journal of Computer Trends and Technology (IJCTT) – volume 8 number 4– Feb 2014, ISSN: 2231-2803, PP 178-183.

[20] Jang and J. R., "Anfis: Adaptive-network-based fuzzy inference systems,"IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, May 1993.

[21] F. Cus, U. Zuperl, M. Milfelner, and B. Mursec, "An adaptive neuro-fuzzy inference system for modeling of end-milling," Faculty of Mechanical Engineering, University ofMaribor, Slovenia, pp. 1-6, 2004.

[22] A. P. Paplinski, "An adaptive neuro-fuzzy inference system," Neuro Fuzzy Computing, pp. 12-1-12-19, May 2005.

Girish Kumar Patnaik received the B.E. degree in Computer Science and Engineering, in 1990 from Marathwada University, Aurangabad. M.E. degree in Computer Science and Engineering from Motilal Nehru Regional College of Engineering, Allahabad, Allahabad university in 2001, and Ph.D. from Motilal Nehru National Institute of Technology Allahabad in 2012. Presently working as Professor and H.O.D. of Computer Engineering department at S.S.B.T. College of Engineering and Technology, Bambhori Jalgaon. (M.S.) having more than 24 years of experience.

Sandip S. Patil received the B.E. degree in Computer Engineering, in 2001 from SSBT's College of Engineering and Technology, Bambhori, Jalgaon affiliated to North Maharastra University Jalgaon(M.S.), M.Tech. degree in Computer Science and Engineering from Samrat Ashok Technological Institute, Vidisha in 2009. Presently working as Associate Professor in department of Computer Engineering at S.S.B.T. College of Engineering and Technology, Bambhori Jalgaon. (M.S.) having more than 12 years of experience.

Archana Neelkanth Mahajan received the B.E. degree in Information technology in 2003 from Godavari College of Engineering , Jalgaon, affiliated to North Maharastra University Jalgaon(M.S.), Pursuing M.E. degree in Computer Science & Engineering from S.S.B.T. College of Engineering and Technology, Bambhori Jalgaon. Presently working as Assistant Professor in department of Computer Engineering at Institute of Technology and Management Universe, Vadodara, Gujarat.