Handling High Dimensionality and Interpretability-Accuracy Trade-Off Issues in Evolutionary Multiobjective Fuzzy Classifiers

Praveen Kumar Shukla, Surya Prakash Tripathi

Abstract—Fuzzy systems are capable to model the inherent uncertainties in real world problems and implement human decision making. In this paper two issues related to fuzzy systems development are addressed and solutions are proposed and implemented. First issue is related to the high dimensional data sets. Such kinds of data sets lead to explode the search space of generated rules and results into deterioration of interpretability and performance of the fuzzy classifiers. To deal with this problem several data clustering algorithms are developed, i.e. fuzzy c-mean clustering algorithm, entropy based fuzzy clustering algorithm etc. The authors have proposed an integrated version of clustering algorithm by replacing the cluster center generation method of fuzzy c-means algorithm by entropy based method. MATLAB is used to implement the proposed fuzzy clustering. On the other hand, interpretability is the subjective feature of fuzzy system that quantifies its understandability. Interpretability can be improved at the cost of other, i.e. improvement in one leads to loss in other. This situation is called 'interpretability-accuracy trade-off'. By handling this trade-off a number of fuzzy systems can be generated with different values of interpretability and accuracy parameters. Evolutionary Multiobjective Optimization is used to deal with this trade-off by proposing a new fuzzy classifier 'Teacher-Performance Fuzzy Classification System'. To implement fuzzy classifier 'Guaje' open access software is used and evolutionary multiobjective optimization framework is implemented using 'MATLAB'.

Index Terms— Fuzzy Classification System, Fuzzy Rule Based Systems, Evolutionary Multiobjective Optimization, Fuzzy c-Means Clustering Algorithm, Interpretability-Accuracy Trade-Off.

1 INTRODUCTION

Fuzzy rule based classifiers [1, 2] are the automatic classification systems in which the knowledge is represented by fuzzy if-then rules. These are the important tools for machine learnining framework.

Designing fuzzy systems for the high dimensional data sets [3] is critical research issue because these data sets leads to exponential growth in terms of rule search space. Several methods are proposed and implemented to develop accurate and interpretable fuzzy systems handling high dimensional data sets. In [4], a method of fuzzy association rule based classification method FARC-HD for high dimensional problems has been proposed and implemented. This approach produces the accurate and compact fuzzy rule based classifiers.

In [5] a high dimensional regression problem has been developed using multiobjective evolutionary algorithms. This framework carries out learning of the database in terms of variables, granularities and displacements in fuzzy partitions.

High dimensional and large data sets are handled in the evolutionary multiobjective framework in [6] also. Multiobjective evolutionary learning of rule base is carried out by selecting reduced set of rules and conditions.

Apart from the issue of high dimensonalty of data sets, interpretability assessment [7,9,12,13] and dealing with interpretability-accuracy trade-off [8,10] are the important research lines. Interpretability has a complete subjective concern to the fuzzy systems. There is no global index to estimate it. Interpretability and accuracy are contradictory with each other. The improvement in interpretability leads to loss in accuracy and this situation is identified as 'interpretability-accuracy trade-off'. A tuning approach has been developed in [11] for interval type-2 fuzzy systems with an improvement in interpretability.

Evolutionary multiobjective optimization algorithms are proved as efficient tools to deal with the situation of trade-off. The interpretability assessment indexes are developed in [14] for evolutionary multiobjective optimization framework.

In this paper, the problem of high dimensional data sets and multiobjective optimization framework for fuzzy classifiers are addressed. In section 2 the fuzzy c-means clustering algorithm is introduced and its improved version is proposed. In section 3 the fuzzy classifier "Engineering Teacher Fuzzy Classification System". Experiments and results are discussed in section 4. Section 5 is the conclusion and future scope.

- *Praveen Kumar Shukla is prsuing Ph. D. in Computer Science & Engineering form Uttar Pradesh Technical University, Lucknow, India. PH-0522-3911310. E-mail:praveenshuklaniec@yahoo.co.in*
- *Surya Prakash Tripathi is Professor and Head in Department of Computer Science at Institute of Engineering & Technology, Lucknow, India. E-mail: tripathee_sp@yahoo.co.in*

2 Feature Selection in High Dimensional Fuzzy Rule Base Classifier (FRBC)

This section introduces the basic concepts of FRBC. The fuzzy c-means clustering algorithm and entropy based clustering algorithm are discussed and a hybrid version of fuzzy c-means clustering algorithm is proposed and further analyzed. .

2.1 Basic concepts of FRBC

FRBC is an automatic classification system in which knowledge is represented by fuzzy rules. KB consists of DB and RB. DB contains Membership Function (MF) and Scaling Function (SF) related to linguistic labels whereas RB consists of set of fuzzy rules. The formal structure of the rule is as follows;

$$if\ x_1\ is\ A_1^m\ and ....and\ x_n\ is\ A_n^m\ then\, Y\ is\, C_i\ with\, d^m$$

Here, $x_1.........x_n$ are the features in the problem and $A_1^m..................A_n^m$ are the linguistic labels which represent the values of linguistic variables. In the consequent part C_i is the class in which the particular object is classified with certainty degree d^m .

Fuzzy inference engine is the function of inferencing classification results as per the knowledge provided by the fuzzy if then rules.

During the design of the FRBC, high dimensionality is the big problem which leads to exponential increment in the number of rules. This deteriorates the interpretability of the system along with its performance and creates possible over-fitting.

To deal with the above problem, two aspects of the FRBC development are considered,

1. Reducing and compacting the rule set leading to minimize the number of fuzzy rules.
2. The selection of appropriate features in the FRBC.

GAs can be well used to deal with the problem of rule reduction and feature selection and it proceeds towards improvement in the interpretability.

2.2 Clustering Algorithms

A cluster is defined as a grouping of similar kind of data point around a center termed as centroid. The precise identification of cluster can be done by ensuring their boundaries. The clusters are identified in two groups; crisp cluster has precise boundaries but fuzzy cluster have fuzzy boundaries. Data clustering is the approach to model different classification problems. This identifies the natural groupings of the data from large data sets to medel the behaviour of the system. Several algorithms have been proposed for fuzzy clustering, like fuzzy c-means clustering algorithm, fuzzy ISODATA, fuzzy k-nearest neighbourhood algorithm etc.

2.2.1 Fuzzy c-Means Clustering Algorithms

Further, fuzzy c-means clustering algorithm [15] is the approach in which each data point belongs to some degree with a specific membership grade. It is used to perform feature selection by reducing the input vector with a very less amount of accuracy loss.

Let $A=\{a_1,a_2,.................,a_n\}$ is a set of given data. The fuzzy c-mean partition of A is a class of fuzzy subsets of a denoted by

$C=\{C_1,C_2,.................,C_n\}$ which satisfies the following conditions

$$\sum_{i=1}^{c} C_i(a_k)=1 \text{ for all } k\in n \text{ and } i=1$$

$$0<\sum_{k=1}^{n} C_i(a_k)\ \ for\ \ all\, i\in n$$

Here c is a positive constant.

Given a set of data $A=\{a_1,a_2,.................,a_n\}$ where a_k in general is a vector of all $k\in n$ .The objective of the fuzzy partition is to find a fuzzy pseudo partition and the associated cluster centres by which the structure of the data is represented as best as possible.

A performance index (PI) is required to be developed and formulated to solve the problem of fuzzy clustering. Normally PI is based on the cluster centres $w_1,w_2,..........,w_c$ associated with the partition approximated by the following formula,

$$w_i=\frac{\sum_{k=1}^{n}[C_i(x_k)]^m x_k}{\sum_{k=1}^{n}[C_i(x_k)]^m} \text{ for all } i\in n$$

Where $m>1$ is a real number that governs the influence of membership grade. The PI of a fuzzy pseudo partition P is denoted by $L_m(p)$ and expressed as follows;

$$L_m(p)=\sum_{k=1}^{n}\sum_{i=1}^{c}[C_i(x_k)]^m\left\|x_k-w_i\right\|^2$$

Here $\left\|x_k-w_i\right\|^2$ represents the distance between x_k and w_i .

2.2.2 Entropy Based Fuzzy Clustering

In this approach the entropy values of data points are calculated and after that the points with minimum entropy values are selected as cluster centers [16].

Follwing assumptions are made for this approach,

1. n dimensional space and m data points are considered to be clustered. Here each data point is represented by x_i where $i=1,2,.......,m$. Now the data set is represented by $m\times n$ matrix.
2. The Euclidian distance between two data points is calculated as below

$$ED_{ij}=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2}$$

3. The similarity between two points is calculated as follows;

$$SI_{ij}=e^{-\eta ED ij}$$

Here η is the numerical constant.

The calculation of η depends on the statement that the similarity value SI_{ij} is equal to 0.5. The distance between two data points become equal to the mean distance MD .

$$MD = \frac{1}{mC_2}\sum_{i=1}^{m}\sum_{j>i}^{m} D_{ij}$$

The calculation of η is done as follows;

$$\eta = \frac{\ln 0.5}{MD}$$

4. Calculation of entropy is done as follows;

$$ENT_i = \sum_{j\in X}^{j\neq i}(SI_{ij}\log_2 SI_{ij}) + (1-SI_{ij})\log_2(1-SI_{ij})$$

2.2.3 Proposed Clustering Approach: Hybrid Entropy Based c-means Fuzzy Clustering

The fuzzy c-mean clustering algorithm is improved by applying the entropy based method to decide the cluster center. The modified algorithm is as follows:

1. n dimensional space and m data points are considered to be clustered. Here each data point is represented by x_i where $i = 1,2,........,m$. Now the data set is represented by $m \times n$ matrix.

2. The Euclidian distance between two data points is calculated as below

$$ED_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$$

3. The similarity between two points is calculated as follows;

$$SI_{ij} = e^{-\eta ED ij}$$

Here η is the numerical constant and will estimated as previous steps.

4. Calculation of entropy is done as follows;

$$ENT_i = \sum_{j\in X}^{j\neq i}(SI_{ij}\log_2 SI_{ij}) + (1-SI_{ij})\log_2(1-SI_{ij})$$

5. The Eucledian distance is calculated between ith data point and the jth cluster center in the pth dimension,

$$D_{ijm} = \| (x_{ip} - CC_{jp} \|)$$

6. Update fuzzy membership matrix U_F as follows:

$$U_{ijp} = \sum_{i=1}^{c}\frac{1}{\left(\frac{D_{ijp}}{D_{icp}}\right)^{\frac{2}{f-1}}}$$

3 EVOLUTIONARY MULTIOBJECTIVE OPTIMIZATION FRAMEWORK: ENGINEERING TEACHER FUZZY CLASSIFICATION SYSTEMS

Multiobjective Evolutionary Algorithms (MOEA) are used to develop fuzzy systems and found efficient in dealing with the interpretability-accuracy trade-off. The basic concepts of evolutionary multiobjective optimization are well discussed in [19,21]. Few frameworks of MOEA for developing fuzzy systems are developed in [17,18]. The review related to evolutionary multiobjective fuzzy systems is carried out in [20,22].

A fuzzy classification system has been proposed to classify the teachers into three classes according to their performance. The input parameters for this decision making are as follows:

1. Teaching quality (Communication, Presentation and Content Delivery) (TQ)
2. Regularity & Punctuality in the Class (RPC)
3. Completion of Syllabus (COS)
4. Student Satisfaction on Quires (SSQ)

According to above input information there are four classifications; 1- Excellent, 2- Good, 3-Average and 4-Poor. The block diagram of fuzzy classifier is given in Fig. 1.

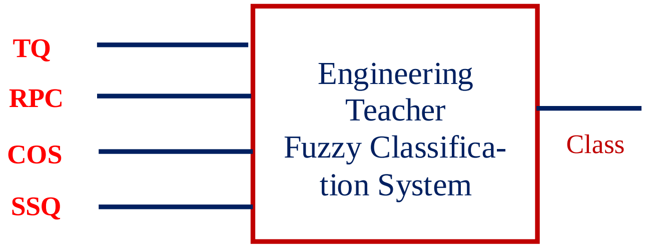

Fig. 1 Engineering Student-Fuzzy Classification System

The membership functions for the above input parameters are as follows:

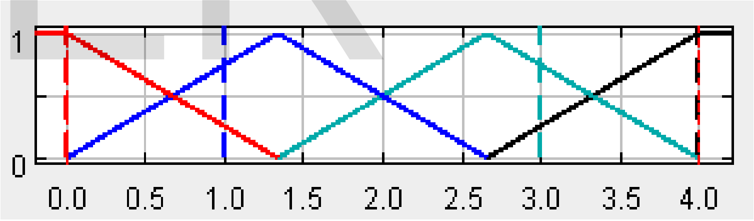

Fig. 2 Membership Function of TQ

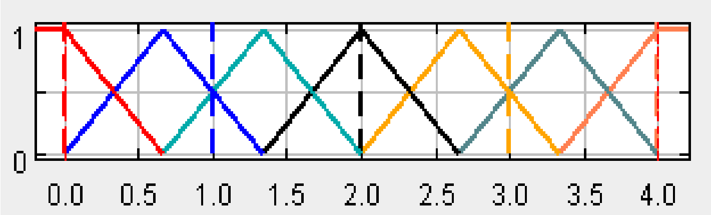

Fig. 3 Membership Function of RPC

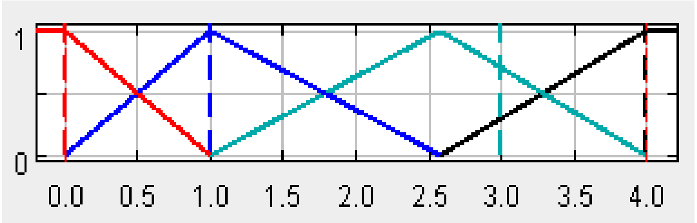

Fig. 4 Membership Function of COS

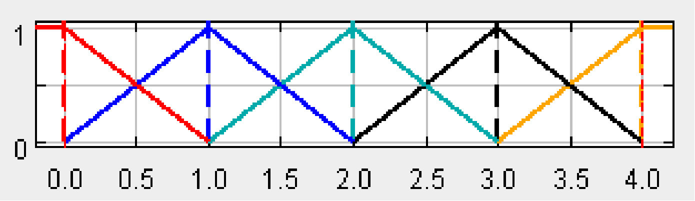

Fig. 5 Membership Function of SSQ

4 EXPERIMENTS & RESULT ANALYSIS

4.1 Feature Selection in High Dimensional Fuzzy Classifiers

To show the performance of proposed approach for feature selection in high dimensional data sets. The experiments are carried out using the data set 'Breast Cancer Wisconsin (Original) Data Set' (Mangasarian and Wolberg (1990)) available in UCI Machine Learning Repository (https://archive.ics.uci.edu/ml) [23].

The attribute information about the data is as follows;

1. Sample code number: id number
2. Clump thickness: 1-10
3. Uniformity of cell size: 1-10
4. Uniformity of cell shape: 1-10
5. Marginal adhesion: 1-10
6. Single epithelial cell size: 1-10
7. Base nuclei: 1-10
8. Bland chromatin: 1-10
9. Normal nucleoli: 1-10
10. Mitoses: 1-10
11. Class (2 for beginning and 4 for malignant)

The other information is given in Table I.

TABLE I
DATA SET CHARACTERISTICS

Characteristics of data sets	**Multivariate**
Number of instances	699
Attribute characteristics	Integer
Number of attributes	10
Type	Classification

The fuzzy rule based classification system has been generated by the open access software 'Guaje' [24].

Experiment-1

The results are summerized in Table II.

Method for rule generation: Wang Mendel

TABLE – II
ACCURACY & INTERPRETABILITY PARAMETERS

Details of Features		**Values**
Accuracy		98.6%
Interpretability	NOR	272
	TRL	2720
	ARL	10
	Average Theoretical Fired Rules	10
	Logical View Index (LVI)	0.657

The clustering results are given in fig. 6. The parameters are set as follows; Maximum iteration: 100, exponent: 2.0, iteration count : 12.

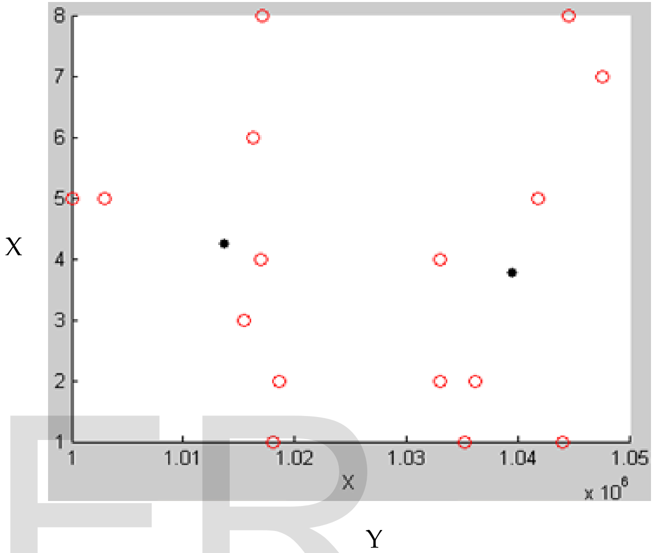

Fig. 6 Cluster Center Identification

Experiment-2

The methods are summarized in Table III.

Method for rule generation: Fuzzy Decision Trees

TABLE - III

Details of Features		**Values**
Accuracy		98.6%
Interpretability	NOR	272
	TRL	2720
	ARL	10
	Average Theoretical Fired Rules	10
	Logical View Index (LVI)	0.657

4.2 Proposed system (Teacher Performance-Fuzzy Classification Systems)

The interpretability of the proposed system is assessed in terms of Number of Rules (NOR), Total Rule Length (TRL), Average Rule Length (ARL), Average Fired Rules (AFR) and Nauck's Index (NI). On the other hand, accuracy is measured in terms of Percentage of Correctly Classified Students (PCS). The results of interpretability and accuracy on different values of NOR are given in Table IV.

TABLE IV

INTERPRETABILITY AND ACCURACY PARAMETERS

Experiments	Interpretability					Accuracy
	Nauck's Index	**NOR**	**TRL**	**ARL**	**AFR**	**PCS$_{tst}$**
E1	0.017	15	60	4	5.26	70.1 %
E2	0.012	20	80	4	7.89	79.9%
E3	0.010	25	100	4	7.22	96.2%

On the different input MFs, the effect of using linguistic hedges has been studied. The linguistic hedges are; more-or-less and strictly. The linguistic modifiers have improved the accuracy as per the results in Table V.

TABLE V

COMPARATIVE RESULTS: LINGUISTIC HEDGES

Condition	Accuracy	Nauck Index	NOR	TRL
Without LH	79.9%	0.012	20	80
With LH	82.3%	0.012	20	80

LH=Linguistic Hedges

As discussed in [25] by Alcala et al. the multiobjective formulations are as follows:

$$g_1(x) = Percentage\ Error\ in\ Correctly\ Classified\ Students\ (PECS)$$

$$g_2(x) = Number\ of\ Rules\ (NOR)$$

$$g_3(x) = Total\ Rule\ Length\ (TRL)$$

Formulation 1:

$$\text{minimize } PECS \text{ and minimize } NOR$$

Formulation 2:

$$\text{minimize } PECS \text{ and minimize } TRL$$

The single objective maximization formulations are as follows:

Formulation 1:

$$f_1(S) = g_1(S) - w_1 g_2(S)$$
$$f_2(S) = g_2(S) - w_1 g_1(S)$$

Formulation 2:

$$f_1(s) = g_1(x) - w_2 g_3(x)$$
$$f_2(s) = g_3(x) - w_2 g_1(x)$$

The nondominated solution as per the formulation 1 and 2 are shown in Fig. 7 and 8.

It is to be noted that the formulations above discussed are indirectly representing the following;

$$\text{minimize error and maximize interpretability}$$

The error is measured in terms of PECS and interpretability is measured in terms of NOR and TRL.

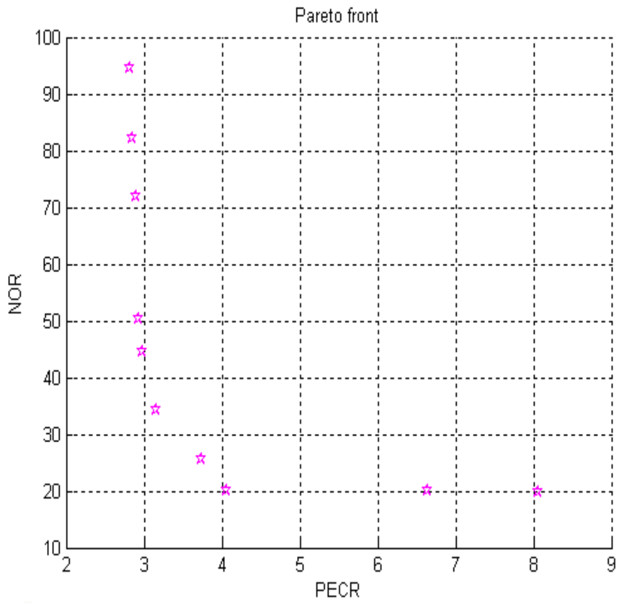

Fig. 7 Pareto Front with Formulation 1

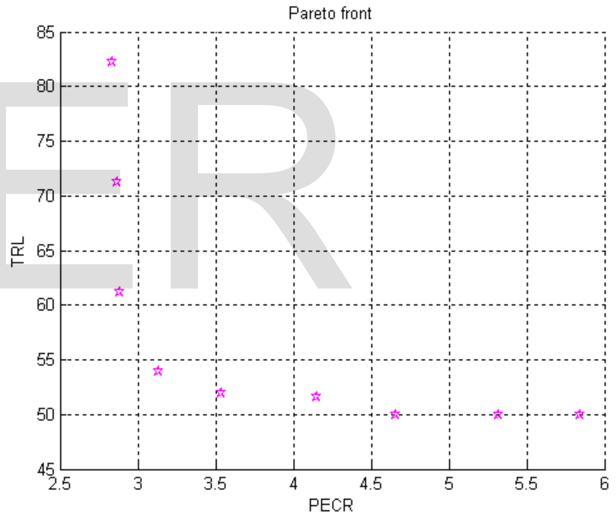

Fig. 8 Pareto Front with Formulation 2

5 Conclusions and Future Scope

High dimensionality of data sets and interpretability accuracy trade-off are two major issues in developing the fuzzy systems. This paper proposes a new approach to deal with high dimensional data sets. The fuzzy c-mean algorithm is improved by replacing its cluster generation method as used in entropy based clustering. The clustering precision of data sets is improved using this approach. The proposal is implemented in MATLAB. Evolutionary multiobjective optimization is used to deal with the interpretability-accuracy trade-off. Multiple fuzzy systems are generated with different trade-off values of interpretability and accuracy parameters. The problem is formulated using multiobjective formulations and Pareto front are generated.

In future the authors will be interested in developing the fuzzy systems dealing with high dimensional data sets in evolutionary multiobjective optimization envioornment. Interpretability improvement and trade-off management would be on the prime concern. The improvement in search capability of evolutionary multiobjective optimization algorithms would also be a new research line particularly in high dimensionality problem.

References

[1] L. Kuncheva, *Fuzzy classifier design*, Berlin: Germany: Springer-Verlag, 2000.

[2] H. Ishibuchi, T. Nakashima, M. Nii, *Classification and modeling with linguistic information granules: advanced approaches to linguistic data mining*, Berlin, Germany: Springer-Verlag, 2000.

[3] Y. Jin, "Fuzzy modeling of high dimensional systems: complexity reduction and interpretability improvement", *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 212-221, 2000.

[4] J. A.-Fdez, R. Alcala, F. Herrera, "A fuzzy association rule based classification model for high dimensional problems with genetic rule selection and lateral tuning", *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857-872, 2011.

[5] M. J. Gacto, R. Alcala, F. Herrera, "Handling high dimensional regression problems by means of an efficient multiobjective evolutionary algorithm", *9th International Conference on Intelligent System Design and Applications*, pp. 109-114, 2009.

[6] M. Antonelli, P. Ducange, F. Marcelloni, "A new approach to handle high dimensional and large data sets in multiobjective evolutionary fuzzy systems", *2011 IEEE International Conference on Fuzzy Systems*, pp. 1286-1293, 2011.

[7] P. K. Shukla, S. P. Tripathi,"Interpretability issues in evolutionary multi-objective fuzzy knowledge base systems", J C Bansal (eds.), *Proceedings of 7th International Conference on Bio-Inspired Computing: Theories and Applications (BICTA-2012), Advances in Intelligent Systems and Computing 201*, pp. 473-484, 2012.

[8] P K Shukla, S. P. Tripathi, "A survey on interpretability-accuracy trade-off in evolutionary fuzzy systems", *2011 5th International Conference on Genetic and Evolutionary Computing*, Taiwan, Xiamen, pp. 97-101, 2011.

[9] P. K. Shukla, S. P. Tripathi, "On the design of interpretable evolutionary fuzzy systems (I-EFS) with improved accuracy", 2012 International Conference on Computing Sciences, pp. 11-14, 2012.

[10] P. K. Shukla, S. P. Tripathi "A review on the Interpretability- Accuracy Trade-Off in Evolutionary Multiobjective Fuzzy Systems (EMOFS)", *Information*, vol. 3, no. 3, pp. 256-277, 2012.

[11] P. K. Shukla, S. P. Tripathi, "A new approach for tuning interval type-2 fuzzy knowledge bases using genetic algorithms", *Journal of Uncertainity Analysis and Applications*, vol. 2, no. 1, pp. 1-15, 2014

[12] J. Cassilas, O. Cordon, F. Herrera, *Interpretability improvements in linguistic fuzzy modelling*, Springer, Heidelberg, Germany, 2003.

[13] M. J. Gacto, R. Alcala, F. Herrera, 'Interpretability of linguistic fuzzy rule based systems: an overview of interpretability measures', *Information Sciences*, Vol. 181, pp. 4340-4360, 2011.

[14] R. Cannone, J. M. Alonso, L. Magdalena, An empirical study on interpretability indexes through Multiobjective evolutionary algorithms, *Springer-Verlag*, Berlin, Heidelberg, Germany, pp. 83-90, 2011.

[15] J. C. Bezdec, Pattern recognition with fuzzy objective function algorithm, *Plenum Press*, New York, 1981.

[16] J. Yao, M. Dash, S. T. Tan, H. Liu, "Entropy based fuzzy clustering and fuzzy modeling", *Fuzzy Sets and Systems*, vol. 113, no. 3, pp. 381-388, 2000.

[17] H. Ishibuchi, Y. Nojima, I. Kuwajima 'Evolutionary Multiobjective design of fuzzy rule based classifiers', *Studies in Computational Intelligence*, vol. 115, pp. 641-685, 2008.

[18] H. Ishibuchi, 'Evolutionary Multiobjective design of fuzzy rule based systems', *2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007)*, Honolulu, HI, USA, pp. 9-16.

[19] K. Deb, *Multiobjective optimization using evolutionary algorithms*, John Wiley & Sons, Chichester, UK, 2001.

[20] H. Ishibuchi, "Multiobjective genetic fuzzy systems: review and future research directions", *2007 FUZZ-IEEE*, London, UK, 23-26 July 2007, pp. 913-918.

[21] P. Ducange, F. Marcelloni *Multiobjective evolutionary fuzzy systems*, Springer-Verlag, Berlin/Heidelberg, Germany, pp. 83-90, 2011.

[22] F. Fazzolari, R. Alcala, Y. Nojima, H. Ishibuchi, F. Herrera, "A review of the application of Multiobjective evolutionary fuzzy systems: current state and further directions", *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 1, pp. 45-65, 2012.

[23] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming", *SIAM News*, vol. 23, No. 5, pp. 1-18, 1990.

[24] J. M. Alonso, L. Magdalena, "HILK++:an interpretability guided fuzzy modelling methodology for learning readable and comprehensible fuzzy rule based classifiers", *Soft Computing*, vol. 15, no. 10, pp. 1959-1980, 2011

[25] R. Alcala, P. Ducange, F. Herrera, B. Lazzerini and F. Marcelloni "A Multiobjective evolutionary approach to concurrently learn rule and data bases of linguistic fuzzy rule-based systems" *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 5, pp. 1106-1122, 2009.