Enhancing Efficiency and Accuracy of Imbalanced Datasets Using Fuzzy Neural Network

S.Lavanya
Department of CSE,
Anna University Regional Centre,
Coimbatore, India.
slavanyamtech@gmail.com

Dr. S. Palaniswami,
Principal,
Government college of Engineering,
Bodinayakanur, India
joegct81@yahoo.com

Abstract— In Data Mining the class Imbalance classification problem is considered to be one of the emergent challenges. This problem occurs when the number of examples that represents one of the classes of the dataset is much lower than the other classes. To tackle with imbalance problem, preprocessing the datasets applied with oversampling method (SMOTE) was previously proposed. Generalized instances are belonging to the family of NGE(abbreviate), which achieves storing objects in Euclidean n-space. The most representative mode used in NGE learning is: classical-BNGE and RISE, recent-INNER, rule induction-RIPPER and PART. In this paper, we propose a Fuzzy Neural Network approach, which is a combination of fuzzy logic and neural networks and called as Neuro Fuzzy System, which could improve the performance and accuracy of the existing system.(explain data set). The proposed approach is compared with NGE learning using SMOTE methods. explain validation/statistical method.

Keywords- Imbalanced Classification, SMOTE, NGE learning, Fuzzy Neural Network, Back propagation

I. INTRODUCTION

Imbalanced Classification problems occur because, the data do not have an equitable distribution among the different classes. This has been recently identified as one of the important problem in Data Mining. Usually, in imbalanced classification problems, the instances are grouped into two types of classes: the majority or negative class, and the minority or positive class. Most learning algorithms obtain a high predictive accuracy over greater number of examples. Furthermore, the examples in the minority can be treated as noise and they might be completely ignored by the classifier. It impacts on several issues namely data loss, dense of data, misclassification problem, over fitting, degrade in accuracy and performance.

Smote-Synthetic Minority Over-sampling Technique [7] is an oversampling approach in which the minority class is over sampled by creating “synthetic” examples rather than by oversampling its replacement. SMOTE provides to form a new minority class examples by interpolating between several minority classes examples that are situated without interruption. Nested Generalized Exemplar (NGE) [3] makes several significant modifications to the exemplar based learning model, this method belonging to the family of the NGE that accomplishes learning by storing objects in Euclidean n-space, which are related to the Nearest Neighbor classifier (NN) [14]. It allows capturing generalizations with exceptions. With respect to instance based classification, the use of generalizations increases the comprehension of the data, reducing the storage requirements. The problem of yielding an optimal number of generalized examples for classifying a set of points is NP-hard. A large but finite subset of them can be easily obtained following a simple heuristic approach over the training data. However, almost all generalized examples produced could be irrelevant and, as a result, the most influential ones must be distinguished.

Fuzzy Neural Network (FNN) or Neuro Fuzzy System (NFS)[5] has been used in imbalanced classification problem with promising results. The combination of fuzzy logic and neural networks is called NFS. They have been successfully used for pattern recognition, regression or density estimation, feature selection, nearest neighbor classification. In this paper, we propose the use of (FNN?)for generalized instances selection in imbalanced classification domains. Our objective is to increase the accuracy of this type of representation by means of selecting the best suitable set of generalized examples to enhance its classification performance over imbalanced domains. We compare our approach with the most representative models of NGE learning: BNGE [14], RISE [13], and INNER [7] and two well-known rule induction learning methods:

RIPPER[15] and PART[12].We have selected a few collections of imbalanced datasets from KEEL-dataset repository [2] for developing our experimental analysis. In order to deal with the problem of imbalanced datasets, preprocessing technique, the SMOTE [7] was used, to balance the distribution of training examples in both classes. The empirical study has been checked via non-parametrical statistical testing, and the results show an improvement of accuracy for our approach whereas the number of generalized examples stored in the final subset was much lower.

Neuro fuzzy system can be discussed as a separate section

Figure 1 shows that the neuro-fuzzy system [19] which attempts to present a fuzzy system in a form of Neural Network [RH99].The neuro-fuzzy system consists of four blocks: fuzzification, multiplication, summation, and division. Fuzzification block translates the input analog signals into fuzzy variables by membership functions. Then, instead of MIN operations in classic fuzzy systems, product operations (signals are multiplied) are performed with fuzzy variables. This neurofuzzy system with product encoding technique is more difficult to implement [OW96], but it can generate a slightly smoother control surface.

The summation and division layers perform defuzzification translation.The weights on upper sum unit are designed as the expecting values (both Mamdani and TSK rules can be used); while the weights on the lower sum unit are all “1”. The defuzzification performs reverse operation of fuzzification. A neural network (NN)[5] is a black box to which we input N values x1,...,xN that form a feature vector x to obtain an output vector z that designates the class, identification, pattern, group or associated output code word of the input vector x. A trained NN represents a system that maps a set of exemplar input feature vectors {x(q): q = 1,...,Q} to a set of output target vectors {t(q): q = 1,...,Q}, also called labels, so that each x(q) maps more closely to t(q) than to another target. This allows the network to make interpolations and extrapolations that map any input x to z that best matches label t (q) for the correct index q.

When trained, a NN is a computational machine that implements an algorithm that is specified by the input nodes, output nodes, layers of hidden nodes between them, connecting lines, functions at the transforming nodes, and weight multipliers w1,...,wR on these lines. We denote the weights here as a vector w. The NN is a composition of many functions designated by the overall NN function as z = f(x;w). Here, x is any input feature vector and z is the resulting output vector that should match best a label t(q) for some q. For this fuzzy neural network there is no adjustment of parameters, no SSEs (sum squared error), no epochs, no explicit rules, no overtraining, and no local or other minima to find.

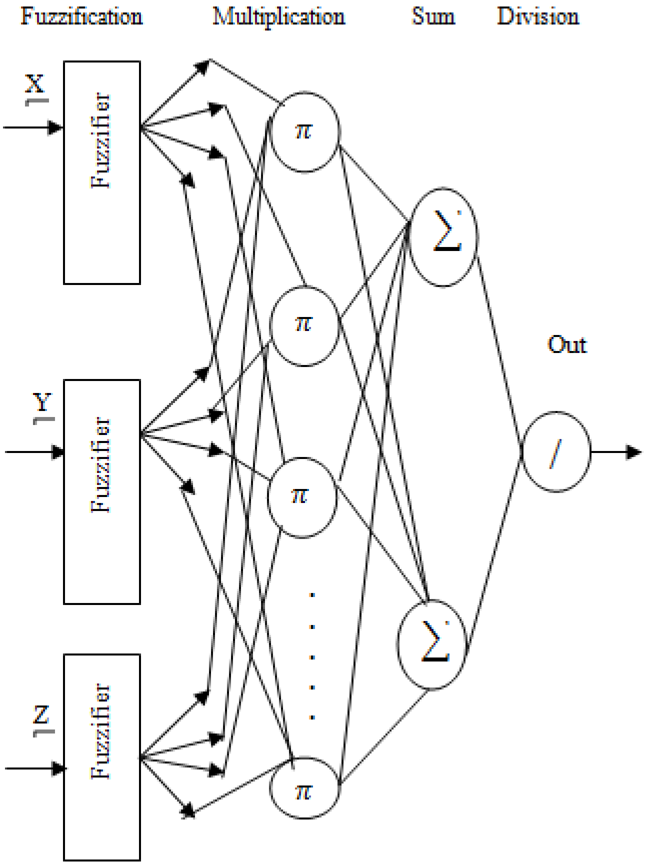


Figure 1: Neuro Fuzzy System

II. ARCHITECTURE

In this system architecture, we describe how the imbalanced datasets are evaluated effectively when compared to the existing method.

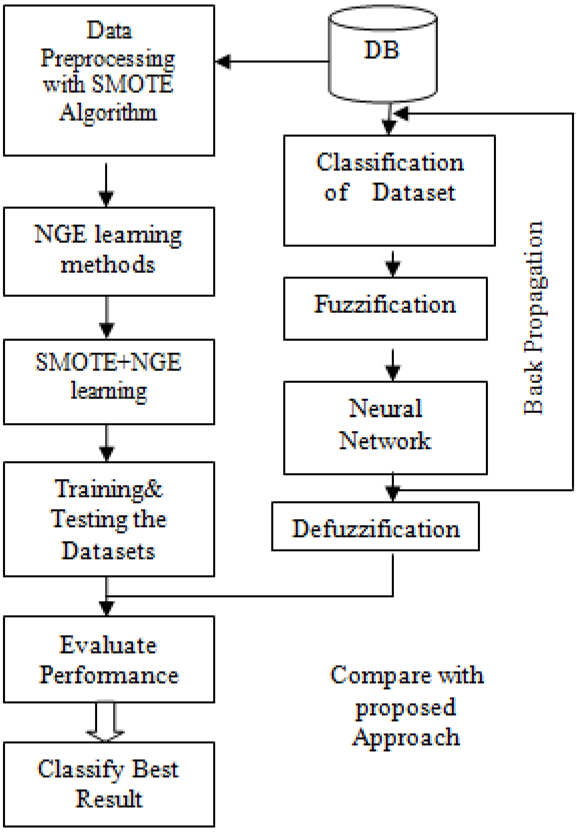


Figure 2: Proposed Architecture

Figure 2 shows how the Fuzzy Neural Network works effectively with imbalanced domains. Data preprocessing with Smote algorithm has been applied to selected datasets from Database. Then NGE learning methods are combined with previous method. In proposed approach of FNN, datasets we've classified among the datasets taken from the DB. The classified datasets are given to the fuzzification [19] where it is used for conversion. The converted input values are passed to the Neural Network. In NN will make decision for imbalanced domains(meaningless sentence). With the help of defuzzification, the fuzzy range values are converted into their original values. In Neural Network, if the imbalanced domains are not satisfied with the existing approaches and then, using back propagation algorithm, the imbalanced datasets are reclassified and similar operations are followed. Both Training and Testing methods are used to measure the performance of existing and proposed system. Finally results are classified from Performance Evaluation.

III. METHODS AND EVALUATION

1. NGE learning

NGE is a learning paradigm based on class exemplars, wherever associate induced hypothesis has the graphical shape of a set of hyper rectangles in an M-dimensional Euclidean space. The input of an NGE system is a set of training examples, each described as a vector of pairs numeric or attribute value and an associated class.

1.1. Matching and Classification:

The matching process is one of the central features in NGE learning and it allows some customization, if desired. Generally this process computes the distance between a new example and an exemplar memory object. The distance is computed as follows (considering numerical attributes):

$$D_{EG} = \sqrt{\sum_{i=1}^{M} \left(\frac{dif_i}{\max_i - \min_i} \right)^2} \qquad (1)$$

$$dif_i = \{E_{fi} - G_{upper} when E_{fi} > G_{upper}$$
$$G_{lower} - E_{fi} when E_{fi} < G_{lower}$$
$$0 - otherwise$$

Where M is the no.of.attributes of the data, E_{fi} is the values of the ith attribute of the example, G_{upper} and G_{lower} are the upper and lower values of G for a specific attribute, max_i and min_i are the maximum and minimum values for ith attribute in training data, respectively. Usually in nominal attributes, the distance is zero when 2 attributes have identical categorical label and 1 on the perverse.

1.2. Suggestions for NGE learning:

BNGE: Batch Nested Generalized Exemplar-BNGE [14] is a batch version of the first model of NGE. The generalization of the examples is done by expanding their boundaries with does not permit overlapping or nesting.

RISE: RISE [13] is an approach proposed to overcome some of the limitations of instance based learning and rule induction by unifying the two. This approach is instance and rule-based induction for unification.

INNER: Inflating examples to obtain rules-INNER [7] starts by selecting a small random subset of examples, which are iteratively inflated in order to cover the surroundings with examples of the same class.

RIPPER: If-Then rules [15] can be extracted directly from the data, but not have to generate a decision tree, using a sequential covering algorithm. Rules are learned for one class at a time.

PART: Generates the rules [12] for selecting the generalized examples in imbalanced classification. It is based on Greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner.

2. Imbalanced Datasets in Classification

We address some important issues related to imbalanced classification by describing the preprocessing technique applied to deal with the imbalance problem: the SMOTE algorithm [7].

2.1. Data preprocessing-SMOTE

Applying a preprocessing step in order to balance the class distribution is a suitable solution to the imbalanced dataset problem. In this approach, the positive class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any of the k minority class nearest neighbors. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbors. This difference could be multiplied by a random number which lie between zero and one, and should added with the feature vector.

2.2. Fuzzy Neural Network (Neuro Fuzzy System)

The main goal of this approach is to ‘fuzzify’ some of the elements of neural networks, using fuzzy logic. In this case, a crisp neuron can become fuzzy. Since FNN are inherently Neural Networks, they're largely employed in Pattern Recog- nition Applications. Both Neural Networks and Fuzzy logic unit [18,19] powerful vogue techniques have their strengths and weaknesses. Neural Networks can learn from datasets, whereas symbolic logic solutions unit straightforward to verify and optimize. Summarizing, Neural Networks can improve their transparency, making them nearer to fuzzy systems. While fuzzy systems will self-adapt, creating them nearer to Neural Networks. For this fuzzy neural network there is no adjustment of parameters (no steepest descent), no SSEs, no epochs, no rules are explicit, There are also no fuzzy rules to learn by training.

2.2.1. Back propagation

Description: (i) Greater convergence speed implying significant reduction [10] in computation time what is important in case of large sized neural networks. (ii) Reaches always forward to the target value without oscillations and there is no possibility to fall into local minimum. (iii) Requires no assumptions about probability distributions and independence of input data and does not require initial weights to be different from each other.

2.2.2. Proposed FNN Algorithm

The high level algorithm is straight forward as given in the steps listed below, and is easy to program. The training and learning reside in the exemplar feature vector and their labels so we don’t need to expend computation time to train. We sometimes use some computation to thin the exemplars that are very similar and belong to the same class. The algorithm [5] is given here in a form for human understanding, but can readily adapt to the program code.

Step 0: Read in the data file (the number of features N, the number of feature vectors Q, the dimension J of the labels, the number K of classes, all Q feature vectors and all Q labels).

```
    Step 1: realize lowest distance Dmin over all feature vector pairs place F = Dmin/2
Put G = letter //Starting no. Gaussian centers
    Step 2: Find two exemplar vectors of min. distance d with indices k1 and k2
If d < (½)Dmin //If vectors are close and
If label [k1] = label [k2] // have same label
Eliminate Gaussian center k2
G = G -1 //Reduce no. Gaussians
Go to Step 2
    Step 3: Input next unknown x to SFNN to be classified,
For k = one to G do //For every Gaussian center
Compute g[k] = exp {-||x - x(k)||2/(2F2)}
Find maximum g[k*], over k = 1,...,G
Output x, label [k*] //label[k*] is class of x
    Step 4: If all inputs for classifying square measure done, stop
Else, go to Step 3.
```

2.3 Evaluation in Imbalanced domains

The most straightforward way to evaluate the performance of classifiers is the analysis based on the confusion matrix. Table 1 illustrates a confusion matrix for a two class problem. Referring the table it is easy to extract wide number of used metrics for performance evaluation of learning systems, such as error rate (2) and accuracy (3).

$$Err = \frac{FP+FN}{TP+FN+FP+TN} \quad (2)$$

$$Acc = \frac{TP+TN}{TP+FN+FP+TN} = 1 - Err \quad (3)$$

Table 1: Confusion matrix for a two-class problem

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

Another appropriate metric that could be used to measure the performance of classification over imbalanced datasets is the Receiver Operating Characteristics (ROC). The Area Under the ROC Curve (AUC) corresponds to the probability of correctly identifying which of the two stimuli is noise and which is signal mixed with noise. AUC provides a summary for the effective performance of learning algorithms. To compute the AUC we just need to obtain the area of the graphics as:

$$\mathrm{AUC} = \frac{1+\mathrm{True}_{\mathrm{Positive}_{\mathrm{Rate}}}-\mathrm{False}_{\mathrm{Positive}_{\mathrm{Rate}}}}{2} \quad (4)$$

True Positive Rate: TP/ (TP+FN) is the percentage of positive cases correctly classified as belonging to the positive class.

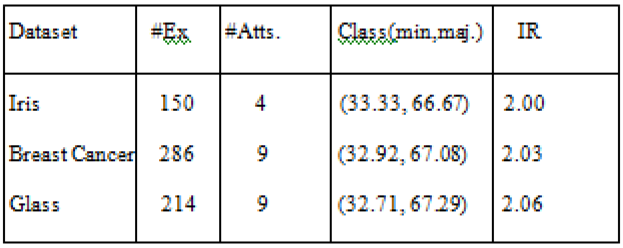
False Positive Rate: FP/ (FP+TN) is the percentage of negative cases misclassified as belonging to the positive class.

IV. EXERIMENTAL FRAMEWORK

1. Data sets and Parameters

This section describes the methodology followed in the experimental study of the generalized examples based learning approaches. We will explain the configuration of the experiment: imbalanced datasets used and parameters for proposed approach.

Table 2: Description for imbalanced datasets



Dataset	#Ex	#Atts.	Class(min,maj.)	IR
Iris	150	4	(33.33, 66.67)	2.00
Breast Cancer	286	9	(32.92, 67.08)	2.03
Glass	214	9	(32.71, 67.29)	2.06

Parameters:

1. Mean vector of the training set;
2. Standard Deviation of the training patterns.

Table 2 summarizes the information chosen during this study and shows, for every knowledge set, the no.of.examples (#Ex), no.of.attributes (#Atts), category name of every class (minority and majority), category attribute distribution and IR?. The table is structured by the IR, as low tounbalanced datasets. The datasets thought of area unit divided exploitation the multiple cross-validations (5-fcv).

We analyze the performance of the methods considering, the entire original without preprocessing datasets. The complete table of results for using algorithms in this study is shown in Table3(i,ii,iii), where the reader can observe the full test results, with their associated standard deviation (SD), in order to compare the performance of each approach. The best case in each dataset is highlighted in bold. We emphasize the good results achieved by FNN, as it obtained the highest AUC value among all algorithms.

Table 3: AUC and SD in test data compare with FNN Approach (in percentage)

(i)

Iris dataset		
Methods	AUC	SD
BNEG	69.4045	1.7842
RISE	76.9548	3.4851
INNER	79.8058	6.2037
RIPPER	85.1313	9.08948
PART	89.3283	12.0168
FNN	93.5845	14.9613

(ii)

Glass dataset		
Methods	AUC	SD
BNEG	69.7802	21.1330
RISE	75.1952	23.6079
INNER	79.1952	29.8248
RIPPER	85.6118	37.9890
PART	89.5240	47.0985
FNN	93.1477	56.6994

(iii)

Breast cancer dataset		
Methods	AUC	SD
BNEG	70.8420	0.8274
RISE	76.0229	3.5661
INNER	79.2000	7.1907
RIPPER	85.9329	10.74542
PART	88.9242	14.3099
FNN	93.7539	17.8766

In this case, the entire table of results with the appliance of the FNN technique is shown in Table two, that achieves the very best end in take a look at among all the algorithms compared during this analysis.

2. Global analysis of results

Finally, we will build a worldwide analysis of outputs by joining the outputs from the tables 3(i,ii,iii). This paper FNN is the best performing one when the datasets are no preprocessed. FNN is a robust algorithm capable to find accurate generalized examples from the original data and it does not require using preprocessed data. Finally, Fuzzy Neural Network was able to see curious behaviour in some NGE learning strategies once they are combined with SMOTE. For this fuzzy neural network there is no adjustment of parameters (no steepest descent), no SSEs(Sum-Squared Error), no epochs, no explicit rules, no overtraining, and no local or other minima to find. There are also no fuzzy rules to learn by training.

Thus the FNN can be used by researchers in other fields who have no experience or intuition about training NNs. The algorithm specifies the computations with no parameters or thresholds to be estimated.

V. RESULTS AND DISCUSSION

Compared with traditional algorithm of NGE learning method, our proposed work using Fuzzy Neural Network is efficient. Fig 3 (Comparing Accuracy) shows that proposed work increases the performance of imbalanced classification when compare with traditional approach. Mainly AUC is used to measure the performance of classification over imbalanced data sets.

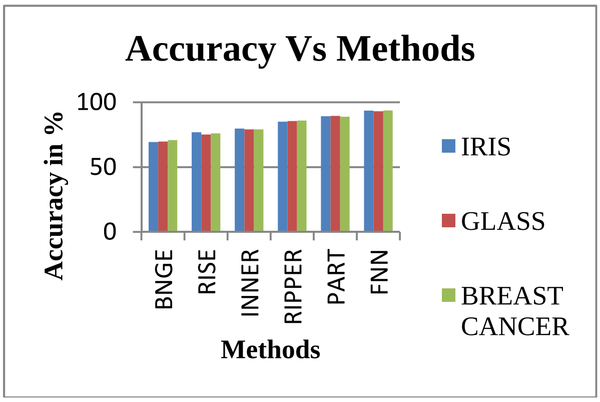


Figure 3: Comparing Accuracy

Figure 4 gives the analysis and comparison of performance offered by NGE learning methods with the proposed method of fuzzy neural network model. Here, if the numbers of layers are increased the standard deviation is increased linearly when compared with traditional approach. Therefore, the best result is find out through higher values of SD. Based on the comparison and also the results from the experiment show the planned approach works higher than existing system.

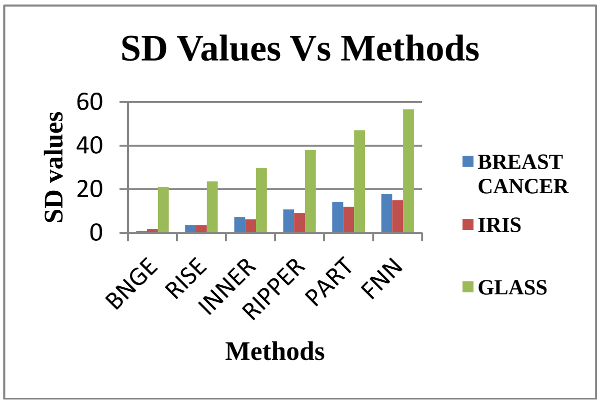


Figure 4: Performance of SD

VI.CONCLUSION AND FUTURE WORK

The purpose of this paper is to present FNN/NFS, an evolutionary model to improve imbalanced classification based on the fuzzy neural based learning. The proposal performs an optimized selection of previously defined generalized examples obtained by a simple and fast heuristic.

The results show that the use of fuzzy neural network algorithm can obtain promising results to optimize the performance in imbalanced domains. It was compared with classical (RISE and BNGE) and recent (INNER) nested generalized learning approaches and two state-of-the-art rule induction methods, RIPPER and PART. EGIS-CHC clearly outperforms all of them when data is not pre-processed. The paper also shows the analysis of using SMOTE as data imbalanced pre-processing and our approach offers dissimilar results in accuracy to the ones offered by the combination of SMOTE with the learning approaches mentioned above, but FNN requires to retain a lower/higher number of generalized examples. In this paper, the imbalanced classification using fuzzy neural network reduces the completion time, error-rate and increases accuracy, performance of imbalanced classification compared with the traditional algorithm.This research introduced a classification model using fuzzy neural network to solve the imbalanced dataset problem with better results. At the same time, it can include a comparative study of other types of neural networks with a different learning algorithm. As a guideline for further research, this model could be extended to include other type’s classification problem with imbalanced dataset.

REFERENCES

[1] Salvador Garcia , Joaquin Derrac , IsaacTriguero, Cristobal J. Carmona , Francisco Herrera , “Evolutionary-based selection of generalized instances for imbalanced classification” ,Knowledge-Based Systems 25 (2012) 3–12.

[2] J.AlcalaFdez,A.Fernandez,J.Derrac,S.Garcia,L.Sanchez,F.Herrara, “Keel data mining software tool: Dataset repository, integration of algorithms and experimental analysis framework”, Journal of multiple valued of and soft computing 17(23)(2011)255-287.

[3] S.Garcia,J.Derrac,J.Luengo,C.J.Carmona, “Evolutionary selection of hyper rectangles in nested generalized exemplar learning”, Applied Soft Computing11(2011)3032-3045.

[4] AlbertoFernandez, MariaJ.del.Jesus,F.Herrara, “On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets”, Expert systems with applications 36(2009)9805-9812.

[5] Carl G.Looney,Sergiu Dascalu, “A simple Fuzzy Neural Network”, University of Nevada Reno,vol.9,No 2,pp.89557,2009.

[6] WenYu, XiaoouLi, “Fuzzy identification using Fuzzy Neural Networks with stable learning algorithms”, IEEE Trans. on Fuzzy Systems, vol 12,No 3,June2004.

[7] O.Luaces, A.Bahamonde, “Inflating examples to obtain rules”, International Journal of Intelligent Systems 18(11)(2003)1113-1143.

[8] N.Chawla,K.Bowyer,I.Hall,W.Kegelmeyar,“SMOTE: Synthetic minority oversampling technique”, Journal Artificial Intelligence Research16(2002)321-357.

[9] Rui-Ping Li,I.Burhen Turksen, “A fuzzy neural network for pattern classification and feature selection”, Fuzzy sets and system130(2002)101-108.

[10] Stefka Stoeva, AlexanderNitov, “A fuzzy backpropagation algorithm”, fuzzy sets and systems112(2000)27-39.

[11] B.Gabrys,A.burgiela “Genaral fuzzy min-max neural networks for clustering and classification”, IEEE Trans.NeuralNetworks, Vol 11,No 3,Pp.769-783,2000.

[12] E.Frank,I.H.Witten, “Generating accurate rule sets without global optimization”,in: Proceedings the Fifteenth International Conference on Machine Learning ,ICMI,1998,pp.144-151.

[13] P.Domingos, “Unifying instance based and rule based induction”, Machine Learning24(1996)141-168.

[14] D.Wettschereck,T.G.Dietterich,“An experimental comparison of the nearest neighbor and nearest hyperrectangle algorithms”, Machine Learning19(1995)5-27.

[15] W.W.cohen, “Fast effective rule induction, in: Proceedings of the Twelfth international conferences on Machine Learning,1995,pp.115-123.

[16] L.A.Zadeh, “Fuzzy Logic, Neural Networks, and Soft computing”.

[17] Rui-ping Li,masao Mukaidono,I.Burhan Turksen, “A Fuzzy Neural Network for Pattern classification and Feature selection”

[18] AjithAbraham,“Neuro Fuzzy Systems :state-of-the-art Modeling Technique”.

[19] HaoYu, Tiantian, BogdanM.Wilamowski, “Neurofuzzysystem”.