International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.9 May - June – 2025
DOI:10.46647/ijetms.2025.v09i03.022 ISSN: 2581-4621

ENHANCING THE EFFECTIVENESS AND ACCURACY OF GENERALIZED INSTANCES OVER IMBALANCED PROBLEM USING ML

Renuga Devi[1], Gomathi [2],Madhu Leha [3] , Madumitha[4]
[1]Assistant professor, Dept. of CSE, Dhanalakshmi Srinivasan University, Trichy,India
[2]Associate Professor,Dept. of CSE,Anna University,BIT campus,Trichy,India
[3]UG Scholar, Dept. of CSE, Dhanalakshmi Srinivasan University, Trichy,India
[4]UG Scholar, Dept. of CSE, Dhanalakshmi Srinivasan University, Trichy,India

Abstract
In Machine Learning (ML), Classification with imbalanced datasets is considered to be a new challenge for researches in the framework of data mining. The imbalance problem occurs in many examples that represents one of the classes of the dataset is much lower than the other classes. To tackle with imbalance problem, pre-processing the datasets applied with oversampling method (SMOTE) was previously proposed. Generalized instances are belonging to the family of Nested Generalized Exemplar, 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. The 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. The proposed approach deals with the comparison of NGE learning without using SMOTE methods.

General Terms: Data pre-processing, cross validation, rule induction.
Keywords: Nested generalized exemplar learning, Imbalanced classification, SMOTE method.

1. INTRODUCTION

An imbalanced class ratio with datasets is a potential challenge in Machine Learning (ML)-based model development systems (Yang & Xu, 2020) Class imbalance occurs when the total number of samples from one class is significantly higher than the other classes (Cui, Jia, Lin, Song, & Belongie, 2019).In both binary and multiclass classification situations, this inequality can be observed (Wang, Dai, Shen, & Xuan, 2021).The data class with the lowest sample is called the minor class, and the data class with the highest sample is called the major class. Major class frequently refers to a negative class in binary classification problems, whereas minor class refers to a positive class. It is currently a significant issue in various domains such as biology, health, finance, telecommunications, and disease diagnosis. As an effect, it is considered one of the most severe problems in data mining (Ahsan, Luna, & Siddique, 2022).

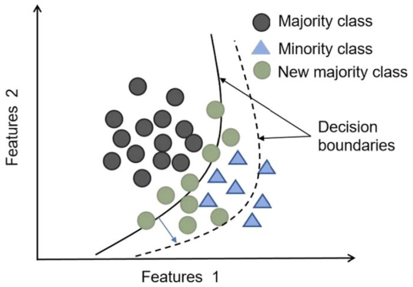


Fig 1 depicts a two-dimensional representation of the major and minor classes.

Data mining is the process of discovering actionable information from large sets of data. Data mining is one of the most important research fields that are due to the expansion of both computer hardware and software technologies, which has imposed organizations to depend heavily on these technologies. Knowledge Discovery in Databases process, or KDD is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract knowledge from a data set in a human-understandable structure.

Classification is the process of finding a model that describe and distinguishes data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or group.

The class imbalance classification problem is one of the current challenges in data mining. It appears when the number of instances of one class is much lower than the instances of the other class. Since standard learning algorithms are developed to minimize the global measure of error, which is independent of the class distribution, in this context this causes a bias towards the majority class in the training of classifiers and results in a lower sensitivity in detecting the minority class examples.

In NGE theory, generalizations take the form of hyperrectangles in an Euclidean n-space. With respect to instance-based classification, the use of generalizations increases the comprehension of the data stored to perform classification of unseen data and the achievement of a substantial compression of the data, reducing the storage requirements. The problem of imbalanced data sets we will include an study that involves the use of a preprocessing technique, the ‘‘Synthetic Minority Over-sampling Technique’’ (SMOTE), to balance the distribution of training examples in both classes.

NGE is a learning paradigm based on class exemplars where an induced hypothesis has the graphical shape of a set of hyperrectangles 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_attribute/value and an associated class. Attributes can either be numerical or categorical. Numerical attributes are usually normalized in the [0, 1] interval Compare the most representative models of NGE learning: BNGE, RISE and INNER and two well-known rule induction learning methods: RIPPER and PART.

BNGE: Batch nested generalized exemplar - The generalization of examples is done by expanding their boundaries just to cover the desired example merging generalized instances only if the new generalized example does not cover (or overlap with) any stored example from a different class. It does not permit overlapping or nesting. RISE: Unifying instance-based and rule-based induction - RISE is an approach proposed to overcome some of the limitations of instance-based learning and rule induction by unifying the two. INNER: Inflating examples to obtain rules - INNER 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.

Fuzzy system based on fuzzy rule based. Fuzzification is supposed to convert the analog inputs into sets of fuzzy variables. For each analog input, several fuzzy variables are generated with values between 0 and 1. The number of fuzzy variables depends on the number of member functions in fuzzification process.

2. RELATED WORK

Our work start with pre-processing the data, estimate accuracy and sd value, evaluate the best method and performance.

2.1 SYSTEM ARCHITECTURE

Figure 1 shows that how the Fuzzy Neural Network works effectively with imbalanced domains. Data Preprocessing is commonly used as a preliminary data mining practice. It transforms the data into a format that will be easily and effectively processed by the users. Data preprocessing with Smote algorithm has been applied to selected datasets from Database.

Smote-Synthetic Minority Over-sampling Technique is associate degree oversampling approach during which the minority category is over sampled by making “synthetic” examples instead of by oversampling with replacement. SMOTE provides to make a replacement minority category examples by interpolating between many minority categories examples that are placed while not interruption.

Nested Generalized example(NGE) makes many vital modifications to the example primarily based learning model, this technique happiness to the family of the NGE that accomplishes learning by storing objects in Euclidian n-space, that are associated with the closest Neighbour classifier(NN). Then NGE learning methods are combined with previous method.

Classify the datasets among taken datasets from DB. The classified datasets are given to the fuzzification where it is used for conversion. The converted input values are passed to the Neural Network.

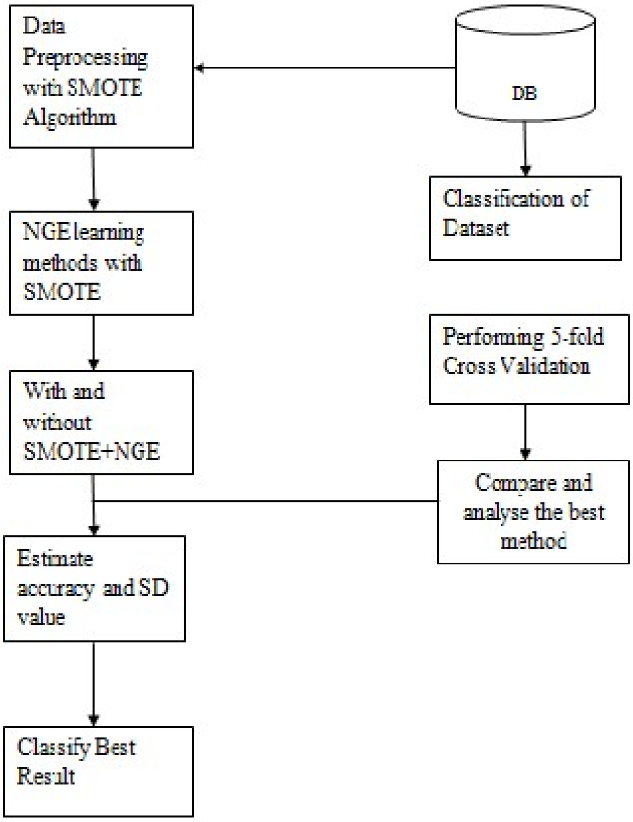

Data Preprocessing with SMOTE Algorithm
DB
Classification of Dataset
NGE learning methods with SMOTE
Performing 5-fold Cross Validation
With and without SMOTE+NGE
Compare and analyse the best method
Estimate accuracy and SD value
Classify Best Result

FIG.1.SYSTEM ARCHITECTURE

In Neural Network will make decision for imbalanced domains. With the help of defuzzification, the fuzzy range values are converted into their original values. Both Training and Testing methods are used to measure the performance of existing and proposed system.

3. METHODS AND EVALUATION

3.1 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.

3.2 Matching and classification

The matching process is one of the central features in NGE learning and this process computes the distance between a new example and an exemplar memory object (a generalized example). This paper, we will refer to the example to be classified as E and the generalized example as G, independently of whether G is formed by a single point or it has some volume. The model computes a match score between E and G by measuring the Euclidean distance between two objects. The Euclidean distance is well-known when G is a single point. The distance is computed as follows (considering numerical attributes):

$$DEG = \sqrt{\sum_{i=1}^{M} \left(\frac{difi}{maxi - mini} \right)^{2}}$$

$$difi = \begin{cases} Efi \_ Gupper \text{ when } Efi > Gupper; \\ Glower \_ Efi \text{ when } Efi < Glower; \\ 0 \text{ otherwise.} \end{cases}$$

Where M is the number of attributes of the data, Efi is the value of the ith attributes of the example, Gupper and Glower are the upper and lower values of G for a specific attribute and maxi and mini are the maximum and minimum values for ith attribute in training data, respectively. Usually in nominal attributes, the distance is zero when two attributes have identical categorical label and one on the perverse.

3.3 Suggestions for NGE learning

BNGE: Batch nested generalized exemplar - The generalization of examples is done by expanding their boundaries just to cover the desired example merging generalized instances only if the new generalized example does not cover (or overlap with) any stored example from a different class. It does not permit overlapping or nesting.

RISE: Unifying instance-based and rule-based induction - RISE is an approach proposed to overcome some of the limitations of instance-based learning and rule induction by unifying the two.

INNER: Inflating examples to obtain rules - INNER 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 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: Generate the ruels 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.

4. IMBALNCED DATASETS IN CLASSIFICATION

4.1 Data pre-processing - SMOTE

The important issues related to imbalanced classification by describing the pre-processing technique applied to deal with the imbalance problem: the SMOTE algorithm. Applying a pre-processing step in order to balance the class distribution is a suitable solution to the imbalanced data set problem.

Data preprocessing refers to remove irrelevant, missing and noisy data. It includes cleaning, normalization transformation, feature extraction and selection. And filling the missing value

In this approach, the positive class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.

A data set contains collection of data. The data set value lists values for each of the variables, such as height and weight of an object for each member of the dataset.

4.2Evaluation of 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 and accuracy.

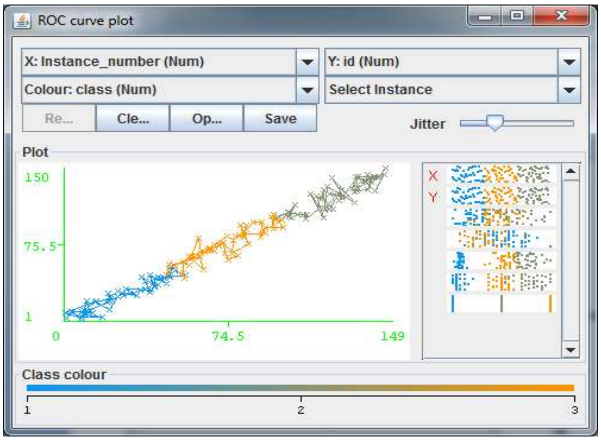
$$Err = \frac{FP+FN}{TP+FN+FP+TN}$$

$$ACC = \frac{TP+TN}{TP+FN+FP+TN} = 1\text{-}Err$$

TABLE 1: CONFUSION MATRIX FOR 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 data sets is the Receiver Operating Characteristic (ROC) graphics. 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 plus noise. AUC provides a single-number summary for the performance of learning algorithms.



ROC curve plot

X: Instance_number (Num)

Y: id (Num)

Colour: class (Num)

Select Instance

Fig: 2 ROC CURVE

The way to build the ROC space is to plot on a two-dimensional chart the true positive rate (Y axis) against the false positive rate (X axis) as shown in Fig 2. The points (0, 0) and (1,1) are trivial classifiers in which the output class is always predicted as negative and positive respectively, while the point (0,1) represents perfect classification. To compute the AUC we just need to obtain the area of the graphic as:

$$AUC = \frac{1+True\_Positive - False\_Positive}{2}$$

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

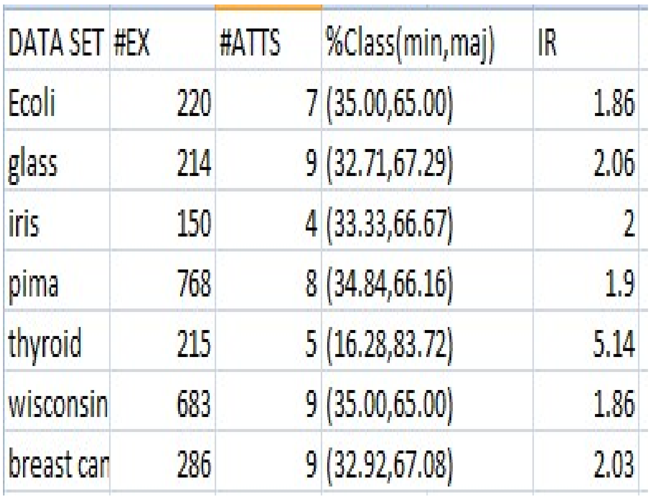
False positive class: FP/(FP+TN) is the percentage of negative class misclassified as belonging to the positive class.

5. EXPERIMENTAL FRAMEWORK

5.1 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



DATA SET	#EX	#ATTS	%Class(min,maj)	IR
Ecoli	220	7	(35.00,65.00)	1.86
glass	214	9	(32.71,67.29)	2.06
iris	150	4	(33.33,66.67)	2
pima	768	8	(34.84,66.16)	1.9
thyroid	215	5	(16.28,83.72)	5.14
wisconsin	683	9	(35.00,65.00)	1.86
breast car	286	9	(32.92,67.08)	2.03

Parameters:

1.Mean vector of the training set
2.Standard deviation of the training patterns.

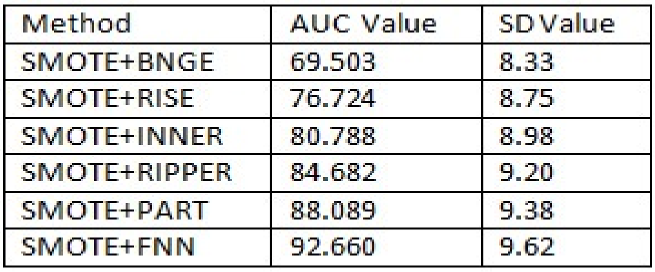
Table 2 summarises the data selected and shows, for each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority), class attribute distribution and IR(imbalance ratios). This table is ordered by the IR, from low to high imbalanced data sets. The data sets considered are partitioned using the fivefold cross-validation (5-fcv) procedure.

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

Table 3(i,ii),where the reader can observe the full test result, with their associated standard deviaton(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.

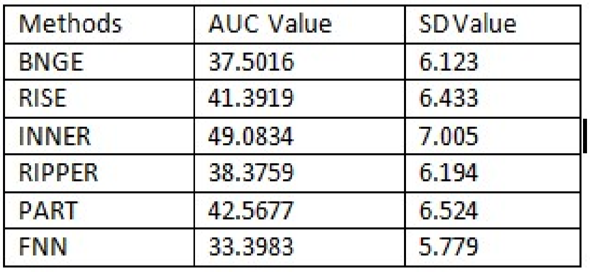
Table 3: AUC and SD in test data compare with FNN approach

(i)IRIS DATASET WITH SMOTE



Method	AUC Value	SD Value
SMOTE+BNGE	69.503	8.33
SMOTE+RISE	76.724	8.75
SMOTE+INNER	80.788	8.98
SMOTE+RIPPER	84.682	9.20
SMOTE+PART	88.089	9.38
SMOTE+FNN	92.660	9.62

(ii)IRIS DATASET WITHOUT SMOTE



Methods	AUC Value	SD Value
BNGE	37.5016	6.123
RISE	41.3919	6.433
INNER	49.0834	7.005
RIPPER	38.3759	6.194
PART	42.5677	6.524
FNN	33.3983	5.779

5.2 Statistical Test

We will use the Wilcoxon signed-rank test as nonparametric statistical procedure for performing pairwise comparisons between two algorithms. For multiple comparisons we use the Friedman test to detect statistical differences among a group of results, and the Holm post hoc test in order to find which algorithms are distinctive among a 1xn comparison. The post hoc procedure allows us to know whether a hypothesis of comparison of means could be rejected at a specified level of significance alpha.

To compute the p-value associated to each comparison, which represents the lowest level of significance of a hypothesis that results in a rejection. It is the adjusted p-value (APV).

5.3 Performance Evaluation

Performance results are measured in terms of the classification accuracy. Classification accuracy is measured in terms of AUC and SD**.** Then compare the Performance evaluation of both existing and proposed system.

5.4 Global analysis of results

Finally, the proposed method, Fuzzy approach, is the best performing one when the data sets are no preprocessed. The combination of SMOTE with FNN produces negative effects by reducing the accuracy in test.

FNN is a robust algorithm capable to find accurate generalized examples from the original data and it does not require to use preprocess data. When the data is treated with SMOTE, an improvement in accuracy is expected. The FNN behaves similarly than other methods combined with SMOTE (1NN, BNGE, INNER, PART and RIPPER), we can emphasize that it requires a lower number of generalized examples or rules than them. Finally, we can see curious behaviors in some NGE learning methods when they are combined with SMOTE.

6. RESULTS AND DISCUSSION

Compared with traditional algorithm of NGE learning method, our proposed work using fuzzy neural network is efficient. Fig 3 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.

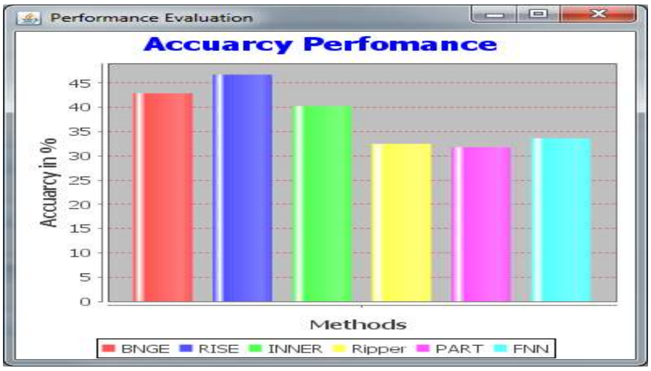


Figure3: comparing accuracy

Figure 4 gives the analysis and comparison of performance offered by NGE learning methods with 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.

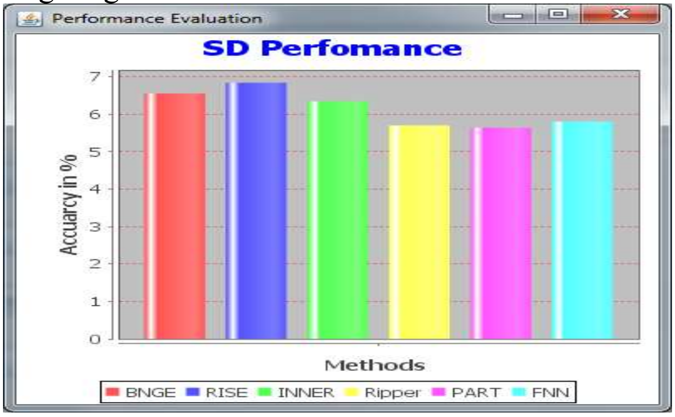


Figure 4:performance of standard deviation

7. CONCLUSION

The purpose of this paper is to present FNN, an evolutionary model to improve imbalanced classification based on the nested generalized example 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 generalized exemplar selection based on evolutionary algorithms 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. The paper also shows the analysis of using SMOTE as data imbalanced preprocessing and our approach offers similar results in accuracy to the ones offered by the combination of SMOTE with the learning approaches mentioned above, but it requires to retain a lower number of generalized examples, thus yielding simpler models.

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.

8. ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher Mrs.s.Lavanya, who gave me the golden opportunity to do project on the topic **"ENHANCING THE EFFECTIVENESS AND ACCURACY OF GENERALIZED INSTANCES OVER IMBALANCED PROBLEM"**, which also helped me in doing a lot of Research and i came to know about so many new things I am really thankful to them.

9. REFERENCES

[1] Salvador Garcia , Joaquin Derrac , Isaac Triguero, Cristobal J. Carmona , Francisc 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] I. Triguero, S. Garcı´ a, F. Herrera, *Differential evolution for optimizing the positioning of prototypes in nearest neighbor classification,* Pattern Recognition 44 (4) (2011) 901–916.

[5] J. Derrac,S. Garcia, F. Herrera, "*IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule",*PatternRecognition43(6)(2010)2082-2105.

[6] A. Fernandez, M.J. del Jesus, F. Herrera, "*On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced datasets",* Information Sciences 180 (8) (2010) 1268–1291.

[7] A. Fernandez, S. Garcia, J. Luengo, E. BernadoMansilla,F.Herrera*,"Geneticsbased machine learning for rule induction: Taxonomy, experimental study and state of the art",* in: IEEE Transactions on Evolutionary Computation 4 (6) (2010) 913–941.

[8] S. Zhang, *Cost-sensitive classification with respect to waiting cost*, Knowledge-Based Systems 23 (5) (2010) 369–378.

[9]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.

[10]Carl G.Looney,Sergiu Dascalu, *"A simple Fuzzy Neural Network",* University of Nevada Reno,vol.9,No 2,pp.89557,(2009).

[11]J. Alcala-Fdez, L. Sanchez, S. Garcı´ a, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas, J. Fernandez, F. Herrera, "*Keel:a software tool to assess evolutionary algorithms for data mining problems",* Soft Computing 13 (3) (2009) 307–318.