International Journal of Artificial Intelligence and Applications for Smart Devices
Vol.2 , No.1 (2014), pp. 23-40
http://dx.doi.org/10.14257/ijaiasd.2014.2.1.03

An Efficient Wrapper approach for Class Imbalance Learning using Intelligent Under-Sampling

Dr. Satuluri Naganjaneyulu[1], Dr. Mrithyumjaya Rao Kuppa[2] and Dr. Ali MIrza Mahmood[3]

[1]*Associate Professor, Lakireddy Bali Reddy College of Engineering, Mylavaram, India*
[2]*Professor, Vaagdevi College of Engineering, Warangal, India.*
[3]*Associate Professor, DMS SVH College of Engineering, Machilipatnam, India.*
svna2198 @gmail.com, mrkuppa @gmail.com, alimirza.md@gmail.com

Abstract

In Data mining and Knowledge Discovery hidden and valuable knowledge from the data sources is discovered. The traditional algorithms used for knowledge discovery are bottle necked due to wide range of data sources availability. Class imbalance is a one of the problem arises due to data source which provide unequal class i.e. examples of one class in a training data set vastly outnumber examples of the other class(es). Researchers have rigorously studied several techniques to alleviate the problem of class imbalance, including resampling algorithms, and feature selection approaches to this problem. In this paper, we present a new hybrid frame work dubbed as Wrapper based Intelligent Under Sampling (WIUS) for learning from skewed training data. These algorithms provide a simpler and faster alternative by using C4.5 and wrapper as base algorithm. We conduct experiments using ten UCI data sets from various application domains using five algorithms for comparison on five evaluation metrics. Experimental results show that our method has higher Area under the ROC Curve, F-measure, precision, TP rate and TN rate values than many existing class imbalance learning methods.

Keywords: *Classification, class imbalance, weighted sampling, WIUS*

1. Introduction

A dataset is class imbalanced if the classification categories are not approximately equally represented. The level of imbalance (ratio of size of the majority class to minority class) can be as huge as 1:99 [1]. It is noteworthy that class imbalance is emerging as an important issue in designing classifiers [2, 3, 4]. Furthermore, the class with the lowest number of instances is usually the class of interest from the point of view of the learning task [5]. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing [6], pollution detection [7], risk management [8], fraud detection [9], and especially medical diagnosis [10–13].

There exist techniques to develop better performing classifiers with imbalanced datasets, which are generally called Class Imbalance Learning (CIL) methods. These methods can be broadly divided into two categories, namely, external methods and internal methods. External methods involve preprocessing of training datasets in order to make them balanced, while internal methods deal with modifications of the learning algorithms in order to reduce their

ISSN: 2288-6710 IJAIASD

sensitiveness to class imbalance [14]. The main advantage of external methods as previously pointed out, is that they are independent of the underlying classifier.

Whenever a class in a classification task is under represented (*i.e.*, has a lower prior probability) compared to other classes, we consider the data as imbalanced [15], [16]. The main problem in imbalanced data is that the majority classes that are represented by large numbers of patterns rule the classifier decision boundaries at the expense of the minority classes that are represented by small numbers of patterns. This leads to high and low accuracies in classifying the majority and minority classes, respectively, which do not necessarily reflect the true difficulty in classifying these classes. Most common solutions to this problem balance the number of patterns in the minority or majority classes. The proposed framework which is shown in Figure 1 addresses the above said issues for class imbalance datasets.

Resampling techniques can be categorized into three groups. Undersampling methods, which create a subset of the original data-set by eliminating instances (usually majority class instances); oversampling methods, which create a superset of the original data-set by replicating some instances or creating new instances from existing ones; and finally, hybrids methods that combine both sampling methods. Among these categories, there exist several different proposals; from this point, we only center our attention in those that have been used in under sampling.

Either way, balancing the data has been found to alleviate the problem of imbalanced data and enhance accuracy [15-17]. Data balancing is performed by, *e.g.*, oversampling patterns of minority classes either randomly or from areas close to the decision boundaries. Interestingly, random oversampling is found comparable to more sophisticated oversampling methods [17]. Alternatively, undersampling is performed on majority classes either randomly or from areas far away from the decision boundaries. We note that random undersampling may remove significant patterns and random oversampling may lead to overfitting, so random sampling should be performed with care. We also note that, usually, oversampling of minority classes is more accurate than undersampling of majority classes [17]. In this paper, we are laying more stress to propose an external CIL method for solving the class imbalance problem.

This paper is organized as follows. Section II briefly reviews the Data Balancing problems and its measures and in Section III, we discuss the proposed method of WIUS (Wrapper based Intelligent Under Sampling) technique for CIL. Section IV presents the imbalanced datasets used to validate the proposed method, while In Section V, we present the experimental setting and In Section VI discuss, in detail, the classification results obtained by the proposed method and compare them with the results obtained by different existing methods and finally, in Section VII we conclude the paper.

2. Literature Review

A comprehensive review of different CIL methods can be found in [18]. The following two sections briefly discuss the external-imbalance and internal-imbalance learning methods. The external methods are independent from the learning algorithm being used, and they involve preprocessing of the training datasets to balance them before training the classifiers. Different resampling methods, such as random and focused oversampling and undersampling, fall into to this category. In random undersampling, the majority-class examples are removed randomly, until a particular class ratio is met [19]. In random oversampling, the minority-class examples are randomly duplicated, until a particular class ratio is met [18]. Synthetic minority oversampling technique (SMOTE) [20] is an oversampling method, where new synthetic examples are generated in the neighborhood of the existing minority-class examples

rather than directly duplicating them. In addition, several informed sampling methods have been introduced in [21].

Currently, the research in class imbalance learning mainly focuses on the integration of imbalance class learning with other AI techniques. How to integrate the class imbalance learning with other new techniques is one of the hottest topics in class imbalance learning research. There are some of the recent research directions for class imbalance learning as follows:

T. Jo *et al.* [22] have proposed a clustering-based sampling method for handling class imbalance problem, while S. Zou *et al.* [23] have proposed a genetic algorithm based sampling method. Jinguha Wang *et al.* [24] have suggested a method for extracting minimum positive and maximum negative features (in terms of absolute value) for imbalanced binary classification is proposed. They have developed two models to yield the feature extractors. Model 1 first generates a set of candidate extractors that can minimize the positive features to be zero, and then chooses the ones among these candidates that can maximize the negative features. Model 2 first generates a set of candidate extractors that can maximize the negative features, and then chooses the ones that can minimize the positive features. Compared with the traditional feature extraction methods and classifiers, the proposed models are less likely affected by the imbalance of the dataset. Iain Brown *et al.* [25] have explored the suitability of gradient boosting, least square support vector machines and random forests for imbalanced credit scoring data sets such as loan default reduction. They progressively increase class imbalance in each of these data sets by randomly undersampling the minority class of defaulters, so as to identify to what extent the predictive power of the respective techniques is adversely affected. They have given the suggestion for applying the random forest and gradient boosting classifiers for better performance. Salvador Garcı´a *et al.* [26] have used evolutionary technique to solve the class imbalance problem. They proposed a method belonging to the family of the nested generalized exemplar that accomplishes learning by storing objects in Euclidean n-space. Classification of new data is performed by computing their distance to the nearest generalized exemplar. The method is optimized by the selection of the most suitable generalized exemplars based on evolutionary algorithms.

Jin Xiao *et al.* [27] have proposed a dynamic classifier ensemble method for imbalanced data (DCEID) by combining ensemble learning with cost-sensitive learning. In this for each test instance, it can adaptively select out the more appropriate one from the two kinds of dynamic ensemble approach: dynamic classifier selection (DCS) and dynamic ensemble selection (DES). Meanwhile, new cost-sensitive selection criteria for DCS and DES are constructed respectively to improve the classification ability for imbalanced data. Victoria López *et al.* [28] have analyzed the performance of data level proposals against algorithm level proposals focusing in cost-sensitive models and versus a hybrid procedure that combines those two approaches. They also lead to a point of discussion about the data intrinsic characteristics of the imbalanced classification problem which will help to follow new paths that can lead to the improvement of current models mainly focusing on class overlap and dataset shift in imbalanced classification. Yang Yong [29] has proposed one kind minority kind of sample sampling method based on the K-means cluster and the genetic algorithm. They used K-means algorithm to cluster and group the minority kind of sample, and in each cluster they use the genetic algorithm to gain the new sample and to carry on the valid confirmation. Chris Seiffert *et al.* [30] have examined a new hybrid sampling/boosting algorithm, called RUSBoost from its individual component AdaBoost and SMOTEBoost, which is another algorithm that combines boosting and data sampling for learning from skewed training data. V. Garcia *et al.* [31] have investigated the influence of both the imbalance ratio and the classifier on the performance of several resampling strategies to deal

with imbalanced data sets. The study focuses on evaluating how learning is affected when different resampling algorithms transform the originally imbalanced data into artificially balanced class distributions.

Table 1 presents recent algorithmic advances in class imbalance learning available in the literature. Obviously, there are many other algorithms which are not included in this table. A profound comparison of the above algorithms and many others can be gathered from the references list.

Table 1. Recent advances in Class Imbalance Learning

ALGORITHM	DESCRIPTION	REFERENECE
DCEID	Combining ensemble learning with cost-sensitive learning.	[37]
RUSBoost	A new hybrid sampling/boosting Algorithm.	[40]
CO2RBFN	A evolutionary cooperative–competitive model for the design of radial-basis function networks which uses both radial-basis function and the evolutionary cooperative-competitive technique.	[42]
Improved FRBCSs	Adapt the 2-tuples based genetic tuning approach to classification problems showing the good synergy between this method and some FRBCSs.	[45]
BSVMs	A model assessment of the interplay between various classification decisions using probability, corresponding decision costs, and quadratic program of optimal margin classifier.	[49]

María Dolores Pérez-Godoy *et al.* [32] have proposed CO2RBFN, a evolutionary cooperative–competitive model for the design of radial-basis function networks which uses both radial-basis function and the evolutionary cooperative-competitive technique on imbalanced domains. CO2RBFN follows the evolutionary cooperative–competitive strategy, where each individual of the population represents an RBF (Gaussian function will be considered as RBF) and the entire population is responsible for the definite solution. This paradigm provides a framework where an individual of the population represents only a part of the solution, competing to survive (since it will be eliminated if its performance is poor) but at the same time cooperating in order to build the whole RBFN, which adequately represents the knowledge about the problem and achieves good generalization for new patterns. Der-Chiang Li *et al.* [33] have suggested a strategy which over-samples the minority class and under-samples the majority one to balance the datasets. For the majority class, they

build up the Gaussian type fuzzy membership function and a-cut to reduce the data size; for the minority class, they used the mega-trend diffusion membership function to generate virtual samples for the class. Furthermore, after balancing the data size of classes, they extended the data attribute dimension into a higher dimension space using classification related information to enhance the classification accuracy. Enhong Che *et al.* [34] have described a unique approach to improve text categorization under class imbalance by exploiting the semantic context in text documents. Specifically, they generate new samples of rare classes (categories with relatively small amount of training data) by using global semantic information of classes represented by probabilistic topic models. In this way, the numbers of samples in different categories can become more balanced and the performance of text categorization can be improved using this transformed data set. Indeed, this method is different from traditional re-sampling methods, which try to balance the number of documents in different classes by re-sampling the documents in rare classes. Such re-sampling methods can cause overfitting. Another benefit of this approach is the effective handling of noisy samples. Since all the new samples are generated by topic models, the impact of noisy samples is dramatically reduced.

Alberto Fernández *et al.* [35] have proposed an improved version of fuzzy rule based classification systems (FRBCSs) in the framework of imbalanced data-sets by means of a tuning step. Specifically, they adapt the 2-tuples based genetic tuning approach to classification problems showing the good synergy between this method and some FRBCSs. The proposed algorithm uses two learning methods in order to generate the RB for the FRBCS. The first one is the method proposed in [36], that they have named the Chi *et al.'s.*, rule generation. The second approach is defined by Ishibuchi and Yamamoto in [37] and it consists of a Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) algorithm.J. Burez *et al.* [38] have investigated how they can better handle class imbalance in churn prediction. Using more appropriate evaluation metrics (AUC, lift), they investigated the increase in performance of sampling (both random and advanced under-sampling) and two specific modeling techniques (gradient boosting and weighted random forests) compared to some standard modeling techniques. They have advised weighted random forests, as a cost-sensitive learner, performs significantly better compared to random forests.

Che-Chang Hsu *et al.* [39] have proposed a method with a model assessment of the interplay between various classification decisions using probability, corresponding decision costs, and quadratic program of optimal margin classifier called: Bayesian Support Vector Machines (BSVMs) learning strategy. The purpose of their learning method is to lead an attractive pragmatic expansion scheme of the Bayesian approach to assess how well it is aligned with the class imbalance problem. In the framework, they did modify in the objects and conditions of primal problem to reproduce an appropriate learning rule for an observation sample. In [40] Alberto Fernández *et al.* have proposed to work with fuzzy rule based classification systems using a preprocessing step in order to deal with the class imbalance. Their aim is to analyze the behavior of fuzzy rule based classification systems in the framework of imbalanced data-sets by means of the application of an adaptive inference system with parametric conjunction operators. Jordan M. Malof *et al.* [41] have empirically investigates how class imbalance in the available set of training cases can impact the performance of the resulting classifier as well as properties of the selected set. In this K-Nearest Neighbor (k-NN) classifier is used which is a well-known classifier and has been used in numerous case-based classification studies of imbalance datasets.

The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake. This skewness towards minority class

(positive) generally causes the generation of a high number of false-negative predictions, which lower the model's performance on the positive class compared with the performance on the negative (majority) class.

3. Wrapper based Intelligent Under Sampling (WIUS)

In this section, we follow a design decomposition approach to systematically analyze the different imbalanced domains. We first briefly introduce the framework design for our proposed algorithm.

The working style of undersampling tries to decrease the number of weak or noise examples. Here, the weak instances related to the specific features are to be eliminated, which is identified according to a well-established wrapper and intelligent technique. The number of instances eliminated will belong to the '*k*' feature selected by wrapper and intelligent technique. Here, the above said routine is employed, which removes examples suffering from feature to class label noises at first and then removes borderline examples and examples of outlier category.

Feature to Class label noises are the examples whose influence is not seen for the decision of the class for that particular feature. Here, they are identified by the limited range categories, using the above said technique. In detail, at first some examples are deleted temporary from N*strong*, a new dataset created with strong instances. Then, for a class to be shrunk, all its examples inside of **N*strong*** are classified. If the classification is correct, and the accuracy is increased then the examples deleted temporary are regarded as being featutre class label noises. Borderline examples are the examples close to the boundaries between different classes for a specific feature. They are unreliable because even a small amount of attribute noise can send the example to the wrong side of the boundary. The outliers are those examples which are very rare in nature from the remaining set of examples. These are examples are of very rare use to the classification and thus to be removed for better performance.

The presented under-sampling algorithm is summarized below.

Algorithm 1 WIUS

Input: A set of minor class examples *P*, a set of major class examples *N*, *jPj* < *jNj*, and *Fj*, the feature set, j > 0.
Output: Average Measure {AUC, Precision, F-Measure, TP Rate, TN Rate }
1: begin
2: $k \leftarrow 0$,j$\leftarrow$1.
3: Apply Wrapper on subset *N*,
4: Find *Fj* from N, k= number of features extracted in CFS
5: repeat
6: k=k+1
7: Select the range for weak or noises instances of *Fj*.
8: Remove ranges of weak attributes and form a set of major class examples N*strong*
9: Until j = k
10: Train and Learn A Base Classifier (C4.5) using *P* and *Nstrong*
11: end

The algorithm 1: WIUS can be explained as follows,

The inputs to the algorithm are minority class “p” and majority class “n” with the number of features j. The output of the algorithm will be the average measures such as AUC, Precision, F-measure, TP rate and TN rate produced by the WIUS method. The algorithm begins with initialization of k=0 and j=1, where k is the number of features extracted by applying correlation based feature subset filter on the dataset and j is the variable used for looping of k features. The ‘k’ value will change from one dataset to other, and depending upon the unique properties of the dataset the value of k can be equal to zero also i.e no attributes can be selected after applying wrapper on the dataset. In this case attributes related under-sampling is not done but overall under-sampling can be performed to remove noise and missing values instances. In any case depending on the amount of majority examples removed, the final "strong set" cannot be balanced i;e number of majority instances and minority instances in the strong set will not be equal.

The different components of our new proposed framework are elaborated in the next subsections.

3.1. Preparation of the Subsets

The datasets is partitioned into majority and minority subsets. As we are concentrating on under sampling, we will take majority data subset for further analysis and reduction.

3.2. Influential Feature Subset Selection

Majority subset can be further analyzed to find the weak or noisy instances so that we can eliminate those. For finding the weak instances one of the ways is that find most influencing attributes or features and then remove ranges of the noisy or weak attributes relating to that feature. How to find the most influencing attribute is by using a wrapper [42], in this case we have used wrapper which uses C4.5 as the algorithm for selecting attributes.

Wrapper technique is proposed by Ron Kohavi [42]. Wrapper is one of the simplest feature selectors conceptually (though not computationally) and has been found to generally out-perform filter methods. Wrapper attribute selection uses the target learning algorithm to estimate the worth of the attribute subsets. Cross-validation is used to provide an estimate for the accuracy of a classifier on novel data when using only the attributes in a given subset. Our implementation uses repeated ten-fold cross validation for accuracy estimation.

3.3. Choosing Feature Class Label and Noise Ranges

How to choose the weak instances relating to that feature from the dataset set? We can find a range where the number of samples are less can give you a simple hint that those instances coming in that range or very rare or noise. We will intelligently detect and remove those instances which are in narrow ranges of that particular feature, borderline and noise instances. The number of features selected by wrapper for each dataset can be reproduced by applying wrapper on the specified datasets. Due to space limitation, we may not able to give all the attributes selected and the ranges of instances removed from the majority subset.

3.4. Forming the Strong Dataset

The minority subset and the stronger majority subset is combined to form a strong and balance dataset, which is used for learning of a base algorithm. In this case we have used C4.5 as the base algorithm.

4. Evaluation Metrics

To assess the classification results we count the number of true positive (TP), true negative (TN), false positive (FP) (actually negative, but classified as positive) and false negative (FN) (actually positive, but classified as negative) examples. It is now well known that error rate is not an appropriate evaluation criterion when there is class imbalance or unequal costs. In this paper, we use AUC, Precision, F-measure, TP Rate and TN Rate as performance evaluation measures.

Let us define a few well known and widely used measures:

The Area under Curve (AUC) measure is computed by equation (1),

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{_____ (1)}$$

The Precision measure is computed by equation (2),

$$\Pr ecision = \frac{TP}{(TP) + (FP)} \quad \text{_____ (2)}$$

The F-measure Value is computed by equation (3),

$$F - measure = \frac{2 \times \Pr ecision \times \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call} \quad \text{_____ (3)}$$

The True Positive Rate measure is computed by equation (4),

$$TruePositiveRate = \frac{TP}{(TP) + (FN)} \quad \text{_____ (4)}$$

The True Negative Rate measure is computed by equation (5),

$$TrueNegativeRate = \frac{TN}{(TN) + (FP)} \quad \text{_____ (5)}$$

5. Experimental Framework

In this section we first describe the collection of imbalanced data sets selected for our study (Section 5.1). Then, we show the algorithms selected for comparison in the experimental study and the corresponding parameters (Section 5.2).

5.1. Evaluation on ten real-world datasets

In this study WIUS is applied to ten binary data sets from the UCI repository [45] with different imbalance ratio (IR). Table 2 summarizes the data selected in this study and shows, for each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and IR.

In order to estimate different measure (AUC, precision, F-measure, TP rate and TN rate) we use a tenfold cross validation approach, that is ten partitions for training and test sets, 90% for training and 10% for testing, where the ten test partitions form the

whole set. For each data set we consider the average results of the ten partitions. We performed the implementation using Weka on Windows XP with 2Duo CPU running on 3.16 GHz PC with 3.25 GB RAM.

Table 2. Summary of Benchmark Imbalanced Datasets

Datasets	# Ex.	# Atts.	Class (_,+)	IR
Breast	268	9	(recurrence; no-recurrence)	1.90
Breast_w	699	9	(benign; malignant)	1.90
Diabetes	768	8	(tested-positive; tested-negative)	1.90
Hepatitis	155	19	(die; live)	1.90
Ionosphere	351	34	(b;g)	2.00
Colic	368	22	(yes,no)	1.90
Vote	435	16	(democrat ; republican)	2.06
Labor	56	16	(bad ; good)	2.06
Sick	3772	29	(negative ; sick)	2.06
Sonar	208	60	(rock ; mine)	2.06

5.2. Algorithms for Comparison and Parameters

To validate the proposed WIUS algorithm, we compared it with the traditional C4.5, CART (Classification and Regression trees), BPN (Back Propagation Neural Networks), REP (Reduced Error Pruning Tree) and SMOTE (Synthetic Minority Oversampling TEchnique). Specifically, we consider five different algorithmic approaches for comparision:

- C4.5: we have selected the C4.5 algorithm as a well-known classifier that has been widely used for imbalanced data. A decision tree consists of internal nodes that specify tests on individual input variables or attributes that split the data into smaller subsets, and a series of leaf nodes assigning a class to each of the observations in the resulting segments. For our study, we chose the popular decision tree classifier C4.5, which builds decision trees using the concept of information entropy. The entropy of a sample S of classified observations is given. C4.5 examines the normalised information gain (entropy difference) that results from choosing an attribute for splitting the data. The attribute with the highest normalised information gain is the one used to make the decision. The algorithm then recurs on the smaller subsets. For this experimental set of C4.5 we have used all the default parameters in WEKA workbench.

- CART: The CART methodology is technically known as binary recursive partitioning. The process is binary because parent nodes are always split into exactly two child nodes and recursive because the process can be repeated by treating each child node as a parent. For this experimental set of CART, we have used all the default parameters in WEKA workbench. The key elements of a CART analysis are a set of rules for:
  i. Splitting each node in a tree;
  ii. Deciding when a tree is complete; and
  iii. Assigning each terminal node to a class outcome (or predicted value for regression).

- BPN: Neural networks (NN) are mathematical representations modeled on the functionality of the human brain. The added benefit of a NN is its flexibility in modeling virtually any non-linear association between input variables and target variable. Although various architectures have been proposed, our study focuses on probably the most widely used type of NN, i.e., the multilayer perceptron (MLP). A MLP is typically composed of an input layer (consisting of neurons for all input variables), a hidden layer (consisting of any number of hidden neurons), and an output layer (in our case, one neuron). Each neuron processes its inputs and transmits its output value to the neurons in the subsequent layer. Each such connection between neurons is assigned a weight during training. During model estimation, the weights of the network are first randomly initialized and then iteratively adjusted so as to minimize an objective function, e.g., the sum of squared errors (possibly accompanied by a regularization term to prevent over-fitting). This iterative procedure can be based on simple gradient descent learning or more sophisticated optimization methods such as Levenberg–Marquardt or Quasi-Newton. we have used one of the algorithm for multilayer perceptron networks design which uses the back propagation algorithm for learning. For this experimental set of BPN, we have used all the default parameters in WEKA workbench.
- REP: One of the simplest forms of pruning is reduced error pruning. Starting at the leaves, each node is replaced with its most popular class. If the prediction accuracy is not affected then the change is kept. While somewhat naive, reduced error pruning has the advantage of simplicity and speed. For this experimental set of REP, we have used all the default parameters in WEKA workbench.
- SMOTE: Regarding the use of the SMOTE pre-processing method [20], we consider only the 1-nearest neighbor (using the euclidean distance) to generate the synthetic samples, and we balance both classes to the 50% distribution. For this experimental set of SMOTE, we have used all the default parameters in WEKA workbench.

6. Results

We evaluated the performance of the proposed WIUS approaches on a number of real-world classification problems. The goal is to examine whether the new proposed learning framework achieve better AUC and other evaluation metrics than a number of existing learning algorithms.

We compared proposed method WIUS with the C4.5, CART, BPN, REP and SMOTE state-of -the-art learning algorithms. In all the experiments we estimate AUC, Precision, F-measure, TP rate and TN rate using 10-fold cross-validation. We experimented with 10 standard datasets for UCI repository; these datasets are standard benchmarks used in the context of high-dimensional imbalance learning. Experiments on these datasets have 2 goals. First, we study the class imbalance properties of the datasets using proposed WIUS learning algorithms. Second, we compare the classification performance of our proposed WIUS algorithm with the traditional and class imbalance learning methods based on all datasets.

Following, we analyze the performance of the method considering the entire original algorithms, without pre-processing, data sets for C4.5, CART, BPN and REP. we also analyze a pre-processing method SMOTE for performance evaluation of WIUS. The complete table of results for all the algorithms used in this study is shown in Table 3 to7, where the reader can observe the full test results, of performance of each approach with their associated standard deviation. We must emphasize the good results achieved by WIUS, as it obtains the highest value among all algorithms

Table 3. Summary of Tenfold Cross Validation Performance for AUC on all the Datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	WIUS
Breast_w	0.957±0.034●	0.950±0.031●	0.991 ±0.018○	0.964±0.038●	0.972±0.027●	0.977±0.021
Diabetes	0.751±0.070●	0.742±0.078●	0.801 ±0.058○	0.751±0.068●	0.792±0.046○	0.759±0.069
Hepatitis	0.668±0.184●	0.561±0.130 ●	0.812 ±0.157○	0.624±0.158●	0.806±0.112○	0.718±0.145
Sonar	0.753±0.113●	0.721±0.106●	0.887 ±0.072○	0.746±0.106●	0.814±0.090○	0.772±0.110
Ionosphere	0.891±0.060●	0.896±0.059●	0.919 ±0.062○	0.902±0.054●	0.904±0.053●	0.914±0.057
Vote	0.979±0.025○	0.973±0.027●	0.985 ±0.013○	0.957±0.023●	0.984±0.017○	0.977±0.030
Colic	0.843±0.070●	0.847±0.070●	0.845 ±0.060●	0.844±0.067●	0.908±0.040○	0.849±0.061
Labor	0.726±0.224●	0.750±0.248●	0.950 ±0.133○	0.767±0.232●	0.833±0.127○	0.804±0.200
Breast	0.606±0.087●	0.587±0.110●	0.645 ±0.109○	0.578±0.116●	0.717±0.084○	0.612±0.109
Sick	0.952±0.040●	0.954±0.043●	0.951 ±0.033●	0.967±0.030○	0.962±0.025●	0.966±0.034

Table 4. Summary of Tenfold Cross Validation Performance for Precision on all the Datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	WIUS
Breast_w	0.965±0.026●	0.971±0.033●	0.976±0.032●	0.962±0.034●	0.976±0.034●	0.979±0.024
Diabetes	0.797±0.045○	0.784±0.041●	0.791±0.053●	0.793±0.044●	0.781±0.062●	0.794±0.053
Hepatitis	0.510±0.371●	0.233±0.337●	0.561±0.308●	0.292±0.391●	0.712±0.175●	0.750±0.393
Sonar	0.728±0.121●	0.709±0.118●	0.822±0.113○	0.733±0.134●	0.863±0.068○	0.804±0.110
Ionosphere	0.895±0.084●	0.868±0.096●	0.952±0.062○	0.886±0.092●	0.934±0.049●	0.947±0.065
Vote	0.971±0.027●	0.971±0.028●	0.959±0.033●	0.969±0.035●	0.977±0.027○	0.972±0.034
Colic	0.851±0.051○	0.853±0.053○	0.851±0.060○	0.857±0.056○	0.853±0.057○	0.844±0.062
Labor	0.696±0.359●	0.715±0.355●	0.867±0.217○	0.698±0.346●	0.871±0.151○	0.782±0.279
Breast	0.753±0.042○	0.728±0.038○	0.763±0.058○	0.721±0.037●	0.710±0.075●	0.723±0.056
Sick	0.992±0.005○	0.992±0.005○	0.980±0.008●	0.990±0.005●	0.983±0.007●	0.991±0.005

Table 5. Summary of Tenfold Cross Validation Performance for F-measure on all the Datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	WIUS
Breast_w	0.962±0.021●	0.960±0.020●	0.973±0.021●	0.963±0.027●	0.961±0.025●	0.978±0.016
Diabetes	0.806±0.044○	0.818±0.045○	0.812±0.420○	0.817±0.045○	0.743±0.058●	0.805±0.041
Hepatitis	0.409±0.272●	0.189±0.231●	0.512±0.257●	0.213±0.267●	0.682±0.149○	0.528±0.301
Sonar	0.716±0.105●	0.672±0.106●	0.800±0.095○	0.689±0.136●	0.861±0.061○	0.761±0.117
Ionosphere	0.850±0.066●	0.841±0.070●	0.859±0.087●	0.848±0.067●	0.905±0.048○	0.895±0.070
Vote	0.972±0.021○	0.966±0.022●	0.954±0.024●	0.961±0.025●	0.969±0.021○	0.967±0.027
Colic	0.888±0.044○	0.890±0.040○	0.849±0.051●	0.882±0.043○	0.880±0.042○	0.879±0.041
Labor	0.636±0.312●	0.660±0.316●	0.861±0.193○	0.650±0.299●	0.793±0.132○	0.749±0.246
Breast	0.838±0.040○	0.813±0.038●	0.764±0.068●	0.805±0.042●	0.730±0.076●	0.823±0.047
Sick	0.993±0.003●	0.994±0.003○	0.984±0.004●	0.993±0.003●	0.987±0.004●	0.994±0.003

Table 6. Summary of Tenfold Cross Validation Performance for TP Rate on all the Datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	WIUS
Breast_w	0.959±0.033●	0.954±0.032●	0.972±0.035●	0.961±0.036●	0.953±0.037●	0.978±0.020
Diabetes	0.821±0.073●	0.852±0.075○	0.842±0.061○	0.841±0.076○	0.712±0.089●	0.825±0.089
Hepatitis	0.374±0.256●	0.172±0.246●	0.523±0.295○	0.192±0.249●	0.681±0.195○	0.438±0.287
Sonar	0.721±0.140●	0.652±0.137●	0.792±0.128○	0.685±0.192●	0.865±0.090○	0.741±0.156
Ionosphere	0.821±0.107●	0.803±0.112●	0.793±0.122●	0.826±0.104●	0.881±0.071○	0.854±0.098
Vote	0.974±0.029○	0.961±0.037●	0.952±0.039●	0.955±0.034●	0.963±0.037●	0.964±0.046
Colic	0.931±0.053○	0.932±0.050○	0.853±0.073●	0.914±0.066●	0.913±0.058●	0.922±0.057
Labor	0.640±0.349●	0.665±0.359●	0.900±0.225○	0.665±0.334●	0.765±0.194●	0.780±0.296
Breast	0.947±0.060●	0.926±0.081●	0.772±0.104●	0.917±0.087●	0.763±0.117●	0.956±0.049
Sick	0.995±0.004●	0.996±0.003○	0.989±0.006●	0.996±0.004○	0.990±0.005●	0.996±0.004

Table 7. Summary of Tenfold Cross Validation Performance for TN Rate on all the Datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	WIUS
Breast_w	0.932±0.052●	0.941±0.056●	0.944±0.062●	0.931±0.068●	0.985±0.028○	0.960±0.047
Diabetes	0.603±0.111●	0.551±0.106●	0.581±0.015●	0.572±0.103●	0.814±0.087○	0.621±0.142
Hepatitis	0.900±0.097●	0.931±0.097●	0.891±0.094●	0.947±0.099●	0.848±0.112●	0.972±0.058
Sonar	0.749±0.134●	0.756±0.121●	0.836±0.122○	0.762±0.145●	0.752±0.113●	0.809±0.123
Ionosphere	0.940±0.055●	0.921±0.066●	0.976±0.030○	0.933±0.063●	0.928±0.057●	0.952±0.062
Vote	0.953±0.045●	0.953±0.046●	0.933±0.057●	0.949±0.059●	0.981±0.023○	0.955±0.057
Colic	0.717±0.119●	0.720±0.114●	0.738±0.118●	0.731±0.121●	0.862±0.063○	0.755±0.109
Labor	0.865±0.197○	0.877±0.192○	0.903±0.159○	0.843±0.214○	0.847±0.187○	0.835±0.214
Breast	0.260±0.141●	0.173±0.164●	0.428±0.160○	0.151±0.164●	0.622±0.137○	0.270±0.164
Sick	0.875±0.071○	0.876±0.078○	0.683±0.123○	0..846±0.080●	0.872±0.053○	0.862±0.070

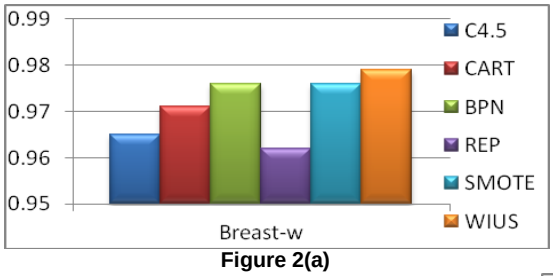

Figure 2(a)

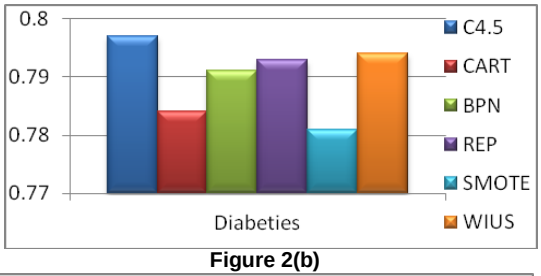

Figure 2(b)

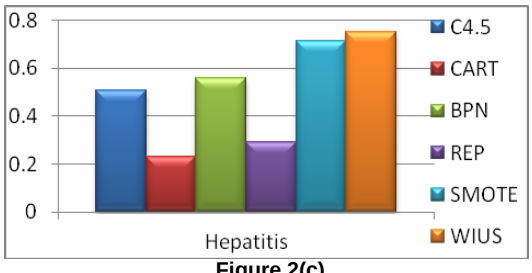

Figure 2(c)

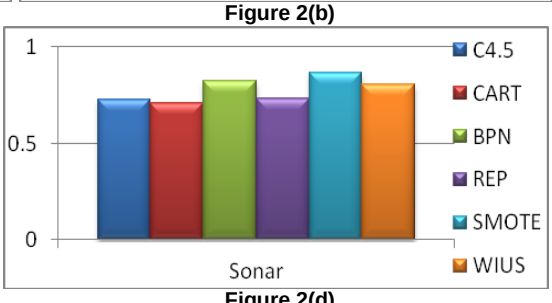

Figure 2(d)

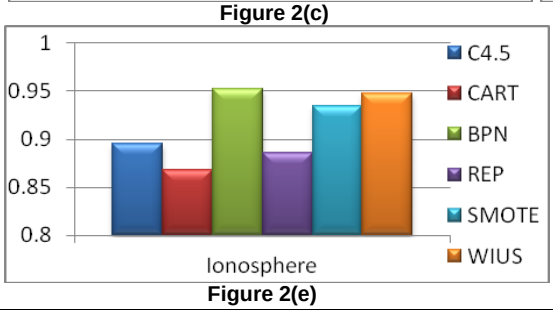

Figure 2(e)

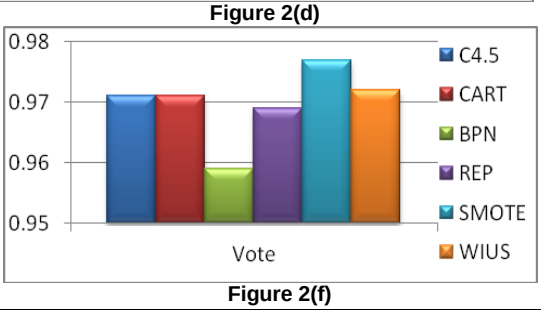

Figure 2(f)

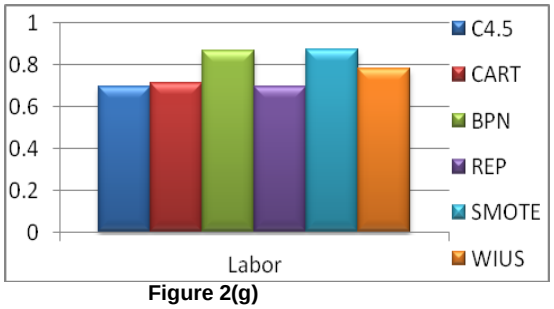

Figure 2(g)

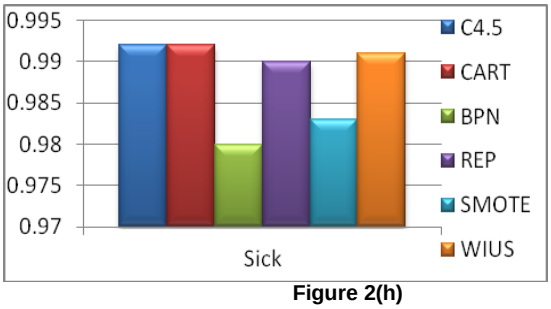

Figure 2(h)

Figure 2(a) – 2(h) Test Results on Precision between the C4.5, CART, BPN, REP, SMOTE and CILIUS-W for Breast_w, Diabeties, Hepatitis, Sonar, Ionosphere, Vote, Labor and Sick Datasets

Figure 2 (a)-(h) shows the average precision computed for all approaches, where we can observe that WIUS has obtained the best precision value in the comparison and therefore it is clearly given the indication of its supremacy. Tables 3, 4, 5, 6 and 7 reports the results of AUC, Precision, F-measure, TP Rate and TN Rate respectively for datasets breast_w, diabetes, hepatitis, sonar, ionosphere, vote, colic, labor, breast and sick. The bold dot '●' indicates a win of WIUS method on C4.5, CART, BPN, REP and SMOTE and a '○' indicates a loss of WIUS method on above said algorithms. The results in the tables show that WIUS has given a good improvement on all the measures of class imbalance learning. This level of analysis is enough for overall projection of advantages and disadvantages of WIUS. A two-tailed corrected resampled paired t-test [46] is used in this paper to determine whether the results of the cross-validation show that there is a difference between the two algorithms is significant or not. Difference in accuracy is considered significant when the p-value is less than 0.05 (confidence level is greater than 95%). In discussion of results, if one algorithm is stated to be better or worse than another then it is significantly better or worse at the 0.05 level.

Finally, we can make a global analysis of results combining the results offered by Tables from 2–7 and Figure 2(a)-(h):

- Our proposals, WIUS is the best performing one when the data sets are no preprocessed. It outperforms the pre-processing SMOTE methods and this hypothesis is confirmed by including standard deviation variations. We have considered a complete competitive set of methods and an improvement of results is expected in the benchmark algorithms i;e C4.5, CART, BPN and REP. However, they are not able to outperform WIUS. In this sense, the competitive edge of WIUS can be seen.
- Considering that WIUS behaves similarly or not effective than SMOTE shows the unique properties of the datasets where there is scope of improvement in minority subset and not in majority subset. Our WIUS can only consider improvements in majority subset which is not effective for some unique property datasets.

Finally, we can say that WIUS is one of the best alternatives to handle class imbalance problems effectively. This experimental study supports the conclusion that the an intelligent under sampling approach with wrapper can improve the CIL behavior when dealing with imbalanced data-sets, as it has helped the WIUS methods to be the best performing algorithms when compared with four classical and well-known algorithms: C4.5, CART, BPN, REP and a well-established pre-processing technique SMOTE.

7. Conclusion

Class imbalance problem have given a scope for a new paradigm of algorithms in data mining. The traditional and benchmark algorithms are worthwhile for discovering hidden knowledge from the data sources, meanwhile Class imbalance Learning methods can improve the results which are very much critical in real world applications. In this paper we present the class imbalance problem paradigm, which exploits the weighted human learning strategy in the supervised learning research area, and implement it with C4.5 and wrapper as its base learners. Experimental results show that WIUS has performed well in the case of multi class imbalance datasets. Furthermore, WIUS is much less volatile than C4.5. In our future work, we will apply WIUS to more learning tasks, especially high dimensional feature learning tasks.

Acknowledgments

We gratefully acknowledge the generous support and valuable suggestion by Associate editor and unanimous reviewers for improving the quality of this work. We would also like to express our thanks to all referees of the paper.

References

[1] J. Wu, S. C. Brubaker, M. D. Mullin and J. M. Rehg, "Fast asymmetric learning for cascade face detection", IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 3, **(2008)** March, pp. 369-382.
[2] N. V. Chawla, N. Japkowicz and A. Kotcz, Eds., Proc. ICML Workshop Learn. Imbalanced Data Sets, **(2003)**.
[3] N. Japkowicz, Ed., Proc. AAAI Workshop Learn. Imbalanced Data Sets, **(2000)**.
[4] G. M.Weiss, "Mining with rarity: A unifying framework", ACM SIGKDD Explor. Newslett., vol. 6, no. 1, **(2004)** June, pp. 7-19.
[5] N. V. Chawla, N. Japkowicz and A. Kolcz, Eds., Special Issue Learning Imbalanced Datasets, SIGKDD Explor. Newsl., vol. 6, no. 1, **(2004)**.
[6] W.-Z. Lu and D. Wang, "Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme", Sci. Total. Enviro., vol. 395, no. 2-3, **(2008)**, pp. 109–116.
[7] Y.-M. Huang, C.-M. Hung and H. C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem", Nonlinear Anal. R. World Appl., vol. 7, no. 4, **(2006)**, pp. 720-747.
[8] D. Cieslak, N. Chawla and A. Striegel, "Combating imbalance in network intrusion datasets", IEEE Int. Conf. Granular Comput., **(2006)**, pp. 732-737.
[9] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Neural Netw., vol. 21, no. 2-3, **(2008)**, pp. 427–436.
[10] A. Freitas, A. Costa-Pereira and P. Brazdil, "Cost-sensitive decision trees applied to medical data," in Data Warehousing Knowl. Discov. (Lecture Notes Series in Computer Science), I. Song, J. Eder, and T. Nguyen, Eds.,
[11] K. Kilic, O. zgeUncu and I. B. Tu¨rksen, "Comparison of different strategies of utilizing fuzzy clustering in structure identification", Inf. Sci., vol. 177, no. 23, **(2007)**, pp. 5153–5162.
[12] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker and R. H. Moss, "A methodological approach to the classification of dermoscopy images," Comput.Med. Imag. Grap., vol. 31, no. 6, **(2007)**, pp. 362-373.
[13] X. Peng and I. King, "Robust BMPM training based on second-order cone programming and its application in medical diagnosis", Neural Netw., vol. 21, no. 2-3, pp. 450–457, 2008.Berlin/Heidelberg, Germany: Springer, (2007), vol. 4654, pp. 303–312.
[14] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning", IEEE Transactions on Fuzzy Systems, vol. 18, no. 3, **(2010)** June, pp. 558-571.
[15] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study", Intelligent Data Analysis, vol. 6, **(2002)**, pp. 429-450.
[16] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection", Proc. 14th Int'l Conf. Machine Learning, **(1997)**, pp. 179-186.

[17] G. E. A. P. A. Batista, R. C. Prati and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data", SIGKDD Explorations, vol. 6, **(2004)** January, pp. 20-29.
[18] D. Cieslak and N. Chawla, "Learning decision trees for unbalanced data", Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer-Verlag, **(2008)**, pp. 241-256.
[19] G. Weiss, "Mining with rarity: A unifying framework", SIGKDD Explor. Newslett., vol. 6, no. 1, **(2004)**, pp. 7-19.
[20] N. Chawla, K. Bowyer and P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", J. Artif. Intell. Res., vol. 16, **(2002)**, pp. 321-357.
[21] J. Zhang and I. Mani, "KNN approach to unbalanced data distributions: A case study involving information extraction", Proc. Int. Conf. Mach. Learning, Workshop: Learning Imbalanced Data Sets, Washington, DC, **(2003)**, pp. 42-48.
[22] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts", ACM SIGKDD Explor. Newslett., vol. 6, no. 1, **(2004)**, pp. 40-49.
[23] S. Zou, Y. Huang, Y. Wang, J. Wang and C. Zhou, "SVM learning from imbalanced data by GA sampling for protein domain prediction", Proc. 9th Int. Conf. Young Comput. Sci., Hunan, China, **(2008)**, pp. 982-987.
[24] J. Wang, J. You, Q. Li and Y. Xu, "Extract minimum positive and maximum negative features for imbalanced binary classification", Pattern Recognition, vol. 45, **(2012)**, pp. 1136-1145.
[25] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications, vol. 39, **(2012)**, pp. 3446–3453.
[26] S. Garcı´a, J. Derrac, I. Triguero, C. J. Carmona and F. Herrera, "Evolutionary-based selection of generalized instances for imbalanced classification", Knowledge-Based Systems, vol. 25, **(2012)**, pp. 3–12.
[27] J. Xiao, L. Xie, C. He and X. Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution", Expert Systems with Applications, vol. 39, **(2012)**, pp. 3668-3675.
[28] V. López, A. Fernández, J. G. Moreno-Torres and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics", Expert Systems with Applications, vol. 39, **(2012)**, pp. 6585–6608.
[29] Y. Yong, "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm", Energy Procedia, vol. 17, **(2012)**, pp. 164-170.
[30] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems And Humans, vol. 40, no. 1, **(2010)** January, pp. 185.
[31] V. Garcia, J. S. Sanchez and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance", Knowledge-Based Systems, vol. 25, **(2012)**, pp. 13-21.
[32] M. Dolores Pérez-Godoy, A. Fernández, A. Jesús Rivera and M. José del Jesus, "Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets", Pattern Recognition Letters, vol. 31, **(2010)**, pp. 2375-2388.
[33] D.-C. Li, C.-W. Liu and S. C. Hu, "A learning method for the class imbalance problem with medical data sets", Computers in Biology and Medicine, vol. 40, **(2010)**, pp. 509-518.
[34] E. Che, Y. Lin, H. Xiong, Q. Luo and H. Ma, "Exploiting probabilistic topic models to improve text categorization under class imbalance", Information Processing and Management, vol. 47, **(2011)**, pp. 202–214.
[35] A. Fernández, M. José del Jesus and F. Herrera, "On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets", Information Sciences, vol. 180, **(2010)**, pp. 1268-1291.
[36] Z. Chi, H. Yan and T. Pham, "Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition", World Scientific, **(1996)**.
[37] H. Ishibuchi, T. Yamamoto and T. Nakashima, "Hybridization of fuzzy GBML approaches for pattern classification problems", IEEE Transactions on System, Man and Cybernetics B, vol. 35, no. 2, **(2005)**, pp. 359-365.
[38] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction", Expert Systems with Applications, vol. 36, **(2009)**, pp. 4626–4636.
[39] C.-C. Hsu, K.-S. Wang and S.-H. Chang, "Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization", Expert Systems with Applications, vol. 38, **(2011)**, pp. 4698-4704.
[40] A. Fernández, M. José del Jesus and F. Herrera, "On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets", Expert Systems with Applications, vol. 36, **(2009)**, pp. 9805–9812.
[41] J. M. Malof, M. A. Mazurowski and G. D. Tourassi, "The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support", Neural Networks, vol. 25, **(2012)**, pp. 141-145.

[42] R. Kohavi, "Wrappers for Performance Enhancement and Oblivious Decision Graphs", PhD thesis, Stanford University, **(1995)**.
[43] J. R. Quinlan, "C4.5: Programs for Machine Learning", 1st ed. San Mateo, CA: Morgan Kaufmann Publishers, **(1993)**.
[44] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representations by back-propagating errors", Nature **323** (6088): 533–536. doi:10.1038/323533a0, **(1986)** October 8.
[45] A. Asuncion and D. Newman, "UCI Repository of Machine Learning Database (School of Information and Computer Science)", Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html, **(2007)**.
[46] C. Nadeau and Y. Bengio, "Inference for the generalization error. Advances in Neural Information Processing Systems", vol. 12, **(2003)**, pp. 307-313.

Authors

Dr. Satuluri Naganjaneyulu, received his Ph.D in Computer Science and Engineering in 2014 from Acharya Nagarjuna University, Guntur, India. M.Tech degree in Computer Science and Engineering from Dr. M.G.R. University, Chennai in 2007. Currently, he is working as an Associate Professor of IT in Lakireddy Bali Reddy College of Engineering, Mylavaram, India. His current research interest includes Data Mining and Knowledge Discovery, Machine Learning, and Artificial Intelligence. He is member of IEEE and Life member of CSI. He has published several papers in reputed Journals.

Dr. Mrithyumjaya Rao Kuppa, received a Ph.D. degree from Kakatiya University in 1979. Now, he is a professor in Faculty of Computer Science and Engineering in Vaagdevi College of engineering, Warangal (India). His current research interest includes data mining techniques with applied to real world problems. He has published in total 16 papers in reputed journals and conferences such as IEEE, Elsevier, Springer and ACM. He also served as a Conference Chair for 2nd Vaagdevi International Conference on Information Technology for real world problems (VCON'10)-2010, http://www.vcon.net.in.

Dr. Ali Mirza Mahmood, received his PhD in Computer Science and Engineering at Acharya Nagarjuna University in 2013. He received his Masters degree in Computer Science in 2003. Now, he is an Associate Professor in department of Computer Science in DMSSVH College of engineering, Machilipatnam (India). His current research interest includes Data Mining and Knowledge Discovery, Machine Learning, and Artificial Intelligence. He has published more than 20 papers in reputed journals and conferences such as IEEE, Elsevier, Springer and ACM. He also served as a Program Committee (PC) Member for 5th International Conference on Information Systems.

