ANALYSIS OF EVALUATION METRICS PERFORMANCE FOR CLASS IMBALANCE CLASSIFICATION PROBLEM

[1]R. Bulli Babu, [2]Dr. Mohammed Ali Hussain

[1]Research Scholar, Department of Computer Science, Bharathair University, Coimbatore, India.
[2]Professor, Department of Electronics & Computer Science Engineering, KLEF, Guntur.

***Abstract*: The class imbalance problem has been recognized in many practical domains and a hot topic of machine learning in recent years. In such a problem, almost all the examples are labeled as one class, while far fewer examples are labeled as the other class, usually the more important class. In this case, standard machine learning algorithms tend to be overwhelmed by the majority class and ignore the minority class since traditional classifiers seeking an accurate performance over a full range of instances. This paper reviewed academic activities special for the class imbalance problem firstly. Then investigated various remedies in four different levels according to learning phases. Following surveying evaluation metrics and some other related factors, this paper showed some future. In this paper, we present a new hybrid frame work and two algorithms dubbed as Class Imbalance Learning using Intelligent Under Sampling—Tree and Neural Network versions (CILIUS-T, CILIUS-NN) for learning from skewed training data. These algorithms pro- vide a simpler and faster alternative by using C4.5 and Neural Network as base algorithm. We conduct experiments using ten UCI datasets from various application domains using five algorithms for comparison on five evaluation metrics. Expert- mental 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**

***Index Terms*— Dataset, class imbalance, Class imbalance ·Weighted sampling · CILIUS**

1. Introduction

Many traditional algorithms to machine learning and data mining problems assume that the target classes share similar prior probabilities. However, in many real world applications, such as oil-spill detection, network intrusion detection, fraud detection, this assumption is grossly violated. In such problems, almost all the examples are labeled as one class, while far fewer examples are labeled as the other class, usually the more important class. This situation is known as the problem of class imbalance.

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 [2–4]. 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 [10–14].

Whenever a class in a classification task is under rep- resented (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. 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, under-sampling is performed on majority classes either randomly or from areas far away from the decision boundaries.

We note that random under-sampling may remove significant patterns and random oversampling may lead to over- fitting, so random sampling should be performed with care. We also note that, usually, oversampling of minority classes is more accurate than under-sampling of majority classes [17]. 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. In this paper the use of an under-sampling technique based on feature selection is proposed in order to create a consistent and strong dataset (without noise), that this method is contrasted with several algorithms and that it is more robust.

This paper is organized as follows: Section 2 briefly presents the literature review conducted for the work and in Sect. 3, we discuss the proposed method of using the Intelligent Under-Sampling (IUS) technique for CIL. Section 4 presents the variations of CILIUS method. Section 5 presents the imbalanced datasets used to validate the pro- posed method, while in Sect. 6, we present the experimental setting and in Sect. 7 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 Sect. 8 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 under-sampling, fall into to this category. In random under-sampling, 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:

Jo et al. [22] have proposed a clustering-based sampling method for handling class imbalance problem, while Zou et al. [23] have proposed a genetic algorithm-based sampling method. 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.

Brown et al. [25] have explored the suitability of gradient boosting, least square support vector machines and random forests for imbalanced credit scoring datasets such as loan default reduction. They progressively increase class imbalance in each of these datasets by randomly under-sampling 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. García et al. [26] have used evolutionary technique to solve the class imbalance problem. They pro- posed 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.

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. 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. 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. 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.

3. Class imbalance learning using intelligent under-sampling (CILIUS)

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 under-sampling 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 filter and intelligent technique. The number of instances eliminated will belong to the *'k'* feature selected by filter 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 shrank, 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 feature 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 proposed framework which is shown in Fig. 1 addresses the problem of class imbalance learning.

The algorithm 1: CILIUS 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 out- put of the algorithm will be the average measures such as AUC, precision, F-measure, TP rate and TN rate produced by the CILIUS 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

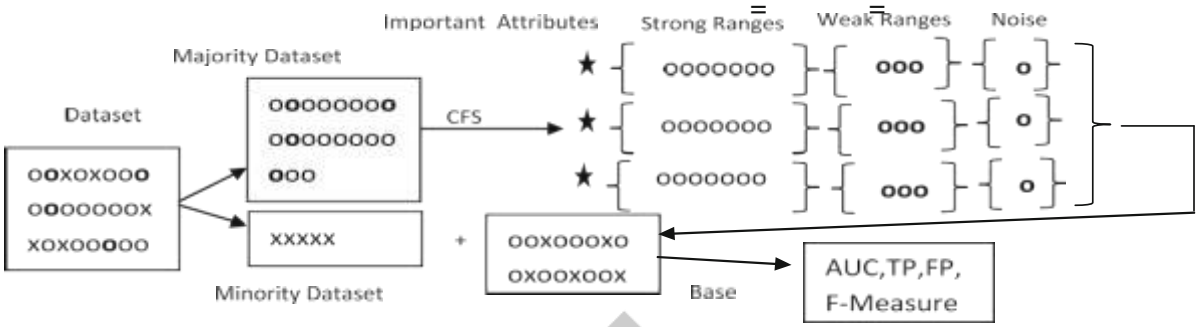


Fig. 1 Framework of class imbalance learning using intelligent under-sampling (CILIUS)

selected after applying CFS, i.e., correlation-based feature subset evaluation [42]. The presented under-sampling algorithm is summarized below:

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← 0, j←1.
3:Apply CFS on subset N,
4:ind Fj from N, k=number of features extracted in CFS
5: repeat
6:k=k+1
7:Select the range for weak noisy instances of Fj
8:Remove ranges of weak attributes and form a set of major class examples Nstrong
9:Until j=k
10:Train and learn a base classifier (C4.5 or NN) using P and Nstrong
11:end

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

Preparation of the subsets :
The datasets are partitioned into majority and minority sub- sets. As we are concentrating on under-sampling, we will take the majority data subset for further analysis and reduction.

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 an attribute selecting filter, in this case we have used CFS [42].

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 in CFS for each dataset can be reproduced by applying CFS 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.

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. Variations of class imbalance learning using intelligent under-sampling (CILIUS)
In the framework of CILIUS a base algorithm is used in the implementation and the efficiency of the CILIUS will also depend on the fine tuning of the parameters and the base algorithms etc. As to find the efficiency of CILIUS for different parameters, we have designed different variations

Table 2 : Summary of bench marked data sets

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	3, 772	29(negative; sick)	2.06
Sonar	208	60(rock; mine)	2.06

Table 3 Summary of tenfold cross-validation performance for AUC on all the datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	CILIUS-T	CILIUS-NN
Breast_w	0.957 ± 0.034	0.950 ± 0.031	0.991 ± 0.018	0.964 ± 0.038	0.972 ± 0.027	0.987 ± 0.016	**0.998 ± 0.003**
Diabetes	0.751 ± 0.070	0.742 ± 0.078	0.801 ± 0.058	0.751 ± 0.068	0.792 ± 0.046	0.826 ± 0.056	**0.878 ± 0.041**
Hepatitis	0.668 ± 0.184	0.561 ± 0.130	0.812 ± 0.157	0.624 ± 0.158	0.806 ± 0.112	0.714 ± 0.166	**0.818 ± 0.135**
Sonar	0.753 ± 0.113	0.721 ± 0.106	0.887 ± 0.072	0.746 ± 0.106	0.814 ± 0.090	0.774 ± 0.114	**0.914 ± 0.059**
Ionosphere	0.891 ± 0.060	0.896 ± 0.059	0.919 ± 0.062	0.902 ± 0.054	0.904 ± 0.053	0.917 ± 0.048	**0.941 ± 0.056**
Vote	0.979 ± 0.025	0.973 ± 0.027	**0.985 ± 0.013**	0.957 ± 0.023	0.984 ± 0.017	0.96 ± 0.034	0.979 ± 0.027
Colic	0.843 ± 0.070	0.847 ± 0.070	0.845 ± 0.060	0.844 ± 0.067	**0.908 ± 0.040**	0.873 ± 0.082	0.875 ± 0.064
Labor	0.726 ± 0.224	0.750 ± 0.248	0.950 ± 0.133	0.767 ± 0.232	0.833 ± 0.127	0.765 ± 0.217	**0.976 ± 0.087**
Breast	0.606 ± 0.087	0.587 ± 0.110	0.645 ± 0.109	0.578 ± 0.116	0.717 ± 0.084	0.637 ± 0.110	**0.878 ± 0.041**
Sick	0.952 ± 0.040	0.954 ± 0.043	0.951 ± 0.033	**0.967 ± 0.030**	0.962 ± 0.025	0.95 ± 0.035	0.941 ± 0.037
Average	0.813 ± 0.090	0.798 ± 0.090	0.879 ± 0.072	0.810 ± 0.089	0.869 ± 0.062	0.842 ± 0.088	0.912 ± 0.055

Bold values indicate highest values of CILIUS by varying the type of base algorithm and fine tuning parameters in implementation. In different variations of CILIUS we deployed C4.5 [43] and Back propagation neural networks [44] as the base algorithm and we called implementations as CILIUS-T and CILIUS-NN respectively

5 . Evaluation metrics

Evaluation criteria :

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 using Eq. (1),

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

The precision measure is computed by using Eqv (2)

$$Precision = \frac{TP}{(TP) + (FP)} \quad (2)$$

The F-measure value is computed by using Eq. (3),

$$F\text{-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

The true positive rate (sensitivity) measure is computed by using Eq. (4),

$$\text{Sensitivity} = \frac{T}{(TP) + (FN)} \quad (4)$$

The true negative rate (specificity) measure is computed by using Eq. (5),

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

Table 4 Summary of tenfold cross-validation performance for precision on all the datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	CILIUS-T	CILIUS-NN
Breast_w	0.965 ± 0.026	0.971 ± 0.033	0.976 ± 0.032	0.962 ± 0.034	0.976 ± 0.034	**0.986 ± 0.020**	0.981 ± 0.023
Diabetes	0.797 ± 0.045	0.784 ± 0.041	0.791 ± 0.053	0.793 ± 0.044	0.781 ± 0.062	0.810 ± 0.048	**0.830 ± 0.048**
Hepatitis	0.510 ± 0.371	0.233 ± 0.337	0.561 ± 0.308	0.292 ± 0.391	**0.712 ± 0.175**	0.698 ± 0.305	0.645 ± 0.283
Sonar	0.728 ± 0.121	0.709 ± 0.118	0.822 ± 0.113	0.733 ± 0.134	0.863 ± 0.068	0.759 ± 0.112	**0.870 ± 0.097**
Ionosphere	0.895 ± 0.084	0.868 ± 0.096	0.952 ± 0.062	0.886 ± 0.092	0.940 ± 0.049	0.922 ± 0.071	**0.943 ± 0.067**
Vote	0.971 ± 0.027	0.971 ± 0.028	0.959 ± 0.033	0.969 ± 0.035	0.977 ± 0.027	**0.978 ± 0.078**	0.957 ± 0.048
Colic	0.851 ± 0.051	0.853 ± 0.053	0.851 ± 0.060	**0.857 ± 0.056**	0.853 ± 0.057	0.787 ± 0.090	0.795 ± 0.094
Labor	0.696 ± 0.359	0.715 ± 0.355	0.867 ± 0.217	0.698 ± 0.346	0.871 ± 0.151	0.754 ± 0.337	**0.983 ± 0.203**
Breast	0.753 ± 0.042	0.728 ± 0.038	0.763 ± 0.058	0.721 ± 0.037	0.710 ± 0.075	0.736 ± 0.050	**0.830 ± 0.048**
Sick	0.992 ± 0.005	**0.992 ± 0.005**	0.980 ± 0.008	0.990 ± 0.005	0.983 ± 0.007	0.990 ± 0.006	0.971 ± 0.011
Average	0.816 ± 0.113	0.784 ± 0.110	0.852 ± 0.094	0.790 ± 0.117	0.811 ± 0.062	0.896 ± 0.112	0.904 ± 0.043

Bold values indicate highest values

6. Experimental framework

Evaluation on ten real-world datasets

In this study CILIUS is applied to ten binary datasets from the UCI repository [45] with different imbalance ratio (IR). Table 2 summarizes the data selected in this study and shows, for each dataset, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and imbalance ratio (IR) [46–49], defined as the ratio of the number of instances of the majority class and the minority class.

In order to estimate different measures (AUC, precision, F-measure, sensitivity and specificity) 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 dataset 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.

Algorithms for comparison and parameters

To validate the proposed CILIUS 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).

Regarding the algorithms for comparison, we have selected alternative paradigms in the CILIUS design field, other neural network models such as multilayer perception, and rule induction algorithms. Specifically, we consider five different algorithmic approaches:

- C4.5: we have selected the C4.5 algorithm as a well- known classifier that has been widely used for imbalanced data. For this experimental set of C4.5 we have used all the default parameters in WEKA workbench.

Default parameters:

Binary Splits = False; Confidence Factor = 0.25; Debug = False; Minimum Number Objects = 2; Number of Folds = 3; Reduced Error Pruning = False; Save Instance Data = False; Seed = 1; Sub tree Raising = True; Unpruned = False; Use Laplace = False.

- CART: The CART methodology is technically known as binary recursive partitioning. For this experimental set of CART, we have used all the default parameters in WEKA workbench.

Default parameters:

Debug = False; Heuristic = True; Minimum Number Objects=2; Number of Folds Pruning = 5; Seed = 1; Size Per = 1.0; Use One

SE = False; Use Prune = True.

- BPN: neural networks (NN) are mathematical represen- tations modeled on the functionality of the human brain. Although various architectures have been proposed, our study focuses on probably the most widely used type of NN, i.e., the multilayer perceptron (MLP). For this experimental set of BPN, we have used all the default parameters in WEKA workbench.

Default parameters:

GUI = False; Auto Build = True; Debug = False; Decay = False; Hidden Layers = a; Learning Rate = 0.3; Momentum = 0.2; Nominal To Binary Filter = True; Nor- malize Attributes = True; Normalize Numeric Class = True; Reset = True; Seed = 0; Training Time = 500; Validation Set Size = 0; Validation Threshold = 20;

Table 5 Summary of tenfold cross-validation performance for F-measure on all the datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	CILIUS-T	CILIUS-NN
Breast_w	0.962 ± 0.021	0.960 ± 0.020	0.973 ± 0.021	0.963 ± 0.027	0.961 ± 0.025	**0.984 ± 0.014**	0.982 ± 0.016
Diabetes	0.806 ± 0.044	0.818 ± 0.045	0.812 ± 0.420	0.817 ± 0.045	0.743 ± 0.058	0.836 ± 0.040	**0.845 ± 0.038**
Hepatitis	0.409 ± 0.272	0.189 ± 0.231	0.512 ± 0.257	0.213 ± 0.267	**0.682 ± 0.149**	0.556 ± 0.238	0.571 ± 0.218
Sonar	0.716 ± 0.105	0.672 ± 0.106	0.800 ± 0.095	0.689 ± 0.136	**0.861 ± 0.061**	0.752 ± 0.103	0.837 ± 0.076
Ionosphere	0.850 ± 0.066	0.841 ± 0.070	0.859 ± 0.087	0.848 ± 0.067	**0.905 ± 0.048**	0.881 ± 0.065	0.865 ± 0.078
Vote	**0.972 ± 0.021**	0.966 ± 0.022	0.954 ± 0.024	0.961 ± 0.025	0.969 ± 0.021	0.963 ± 0.063	0.939 ± 0.044
Colic	0.888 ± 0.044	**0.890 ± 0.040**	0.849 ± 0.051	0.882 ± 0.043	0.880 ± 0.042	0.827 ± 0.073	0.792 ± 0.074
Labor	0.636 ± 0.312	0.660 ± 0.316	0.861 ± 0.193	0.650 ± 0.299	0.793 ± 0.132	0.697 ± 0.307	**0.870 ± 0.191**
Breast	0.838 ± 0.040	0.813 ± 0.038	0.764 ± 0.068	0.805 ± 0.042	0.730 ± 0.076	0.812 ± 0.046	**0.845 ± 0.038**
Sick	0.993 ± 0.003	**0.994 ± 0.003**	0.984 ± 0.004	0.993 ± 0.003	0.987 ± 0.004	0.991 ± 0.004	0.976 ± 0.007
Average	0.743 ± 0.004	0.761 ± 0.043	0.816 ± 0.006	0.798 ± 0.004	0.765 ± 0.042	0.852 ± 0.006	0.872 ± 0.012

Bold values indicate highest values

Table 6 Summary of tenfold cross-validation performance for sensitivity on all the datasets

Datasets	C4.5	CART	BPN	REP	SMOTE	CILIUS-T	CILIUS-NN
Breast_w	0.959 ± 0.033	0.954 ± 0.032	0.972 ± 0.035	0.961 ± 0.036	0.953 ± 0.037	**0.984 ± 0.022**	0.984 ± 0.021
Diabetes	0.821 ± 0.073	0.852 ± 0.075	0.842 ± 0.061	0.841 ± 0.076	0.712 ± 0.089	**0.869 ± 0.064**	0.864 ± 0.063
Hepatitis	0.374 ± 0.256	0.172 ± 0.246	0.523 ± 0.295	0.192 ± 0.249	**0.681 ± 0.195**	0.499 ± 0.525	0.581 ± 0.272
Sonar	0.721 ± 0.140	0.652 ± 0.137	0.792 ± 0.128	0.685 ± 0.192	**0.865 ± 0.090**	0.762 ± 0.114	0.822 ± 0.115
Ionosphere	0.821 ± 0.107	0.803 ± 0.112	0.793 ± 0.122	0.826 ± 0.104	**0.881 ± 0.071**	0.853 ± 0.103	0.809 ± 0.119
Vote	**0.974 ± 0.029**	0.961 ± 0.037	0.952 ± 0.039	0.955 ± 0.034	0.963 ± 0.037	0.951 ± 0.062	0.926 ± 0.074
Colic	0.931 ± 0.053	**0.932 ± 0.050**	0.853 ± 0.073	0.914 ± 0.066	0.913 ± 0.058	0.884 ± 0.113	0.803 ± 0.111
Labor	0.640 ± 0.349	0.665 ± 0.359	**0.900 ± 0.225**	0.665 ± 0.334	0.765 ± 0.194	0.690 ± 0.332	0.890 ± 0.231
Breast	**0.947 ± 0.060**	0.926 ± 0.081	0.772 ± 0.104	0.917 ± 0.087	0.763 ± 0.117	0.909 ± 0.071	0.864 ± 0.063
Sick	0.995 ± 0.004	0.996 ± 0.003	0.989 ± 0.006	**0.996 ± 0.004**	0.990 ± 0.005	0.992 ± 0.006	0.982 ± 0.012
Average	0.874 ± 0.004	0.856 ± 0.123	0.886 ± 0.036	0.871 ± 0.104	0.912 ± 0.012	0.905 ± 0.132	0.896 ± 0.023

Bold values indicate highest values

- REP: one of the simplest forms of pruning is reduced error pruning. For this experimental set of REP, we have used all the default parameters in WEKA workbench.

Default parameters:

Debug = False; Max Depth = −1; Minimum Numbers = 2.0; Minimum Variance Proposition=0.0010; No Pruning = False; Number of Folds = 3; Seed = 1.

- 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 % dis- tribution. For this experimental set of SMOTE, we have used all the default parameters in WEKA workbench.

Default parameters:

Class Value = 0; Nearest Neighbors = 5; Percentage = 100.0; Random Seed = 1.

Table 7 Summary of tenfold cross-validation performance for specificity on all the datasets

Datasets	C4.5	CART	BPN	REP	MOTE	CILIUS-T	CILIUS-NN
Breast_w	0.932 ± 0.052	0.941 ± 0.056	0.944 ± 0.062	0.931 ± 0.068	**0.985 ± 0.028**	0.978 ± 0.030	0.971 ± 0.036
Diabetes	0.603 ± 0.111	0.551 ± 0.106	0.581 ± 0.015	0.572 ± 0.103	**0.814 ± 0.087**	0.696 ± 0.096	0.734 ± 0.094
Hepatitis	0.900 ± 0.097	0.931 ± 0.097	0.891 ± 0.094	**0.947 ± 0.099**	0.848 ± 0.112	0.920 ± 0.092	0.868 ± 0.121
Sonar	0.749 ± 0.134	0.756 ± 0.121	0.836 ± 0.122	0.762 ± 0.145	0.752 ± 0.113	0.743 ± 0.138	**0.864 ± 0.118**
Ionosphere	0.940 ± 0.055	0.921 ± 0.066	**0.976 ± 0.030**	0.933 ± 0.063	0.928 ± 0.057	0.948 ± 0.052	0.965 ± 0.042
Vote	0.953 ± 0.045	0.953 ± 0.046	0.933 ± 0.057	0.949 ± 0.059	0.981 ± 0.023	**0.983 ± 0.030**	0.967 ± 0.037
Colic	0.717 ± 0.119	0.720 ± 0.114	0.738 ± 0.118	0.731 ± 0.121	**0.862 ± 0.063**	0.765 ± 0.122	0.793 ± 0.120
Labor	0.865 ± 0.197	0.877 ± 0.192	0.903 ± 0.159	0.843 ± 0.214	0.847 ± 0.187	0.865 ± 0.207	**0.918 ± 0.152**
Breast	0.260 ± 0.141	0.173 ± 0.164	0.428 ± 0.160	0.151 ± 0.164	0.622 ± 0.137	0.325 ± 0.156	**0.734 ± 0.094**
Sick	0.875 ± 0.071	0.876 ± 0.078	0.683 ± 0.123	0.846 ± 0.080	0.872 ± 0.053	**0.903 ± 0.060**	0.708 ± 0.112
Average	0.809 ± 0.004	0.813 ± 0.003	0.843 ± 0.006	0.832 ± 0.004	0.890 ± 0.015	0.885 ± 0.006	0.876 ± 0.012

Bold values indicate highest values

Table 8 Summary of experimental results for some of the datasets using non-parametric Wilcoxon test

	Breast_w		Diabetes		Hepatitis		Sonar		Ionosphere	
	CILIUS-T	CILIUS-NN	CILIUS-T	CILIUS-NN	CILIUS-T	CILIUS-NN	CILIUS-T	CILIUS-NN	CILIUS-T	CILIUS-NN
AUC	4/5	5/5	5/5	5/5	3/5	5/5	3/5	5/5	4/5	5/5
Precision	5/5	5/5	5/5	5/5	4/5	4/5	3/5	5/5	3/5	4/5
F-measure	5/5	5/5	5/5	5/5	4/5	4/5	3/5	4/5	4/5	4/5
Sensitivity	5/5	5/5	5/5	5/5	3/5	3/5	3/5	4/5	4/5	2/5
Specificity	1/5	1/5	1/5	1/5	3/5	4/5	5/5	0/5	1/5	1/5

Table 9 Summary of experimental results some of the datasets using non-parametric Wilcoxon test

Vote			Colic		Labor		Breast_Cancer		Sick	
	CILIUS-T	CILIUS-NN	CILIUS-T	ILIUS-NN	CILIUS-T	CILIUS-NN	CILIUS-T	CILIUS-NN	CILIUS-T	CILIUS-NN
AUC	1/5	3/5	4/5	4/5	2/5	5/5	3/5	5/5	3/5	0/5
Precision	5/5	1/5	0/5	0/5	3/5	5/5	3/5	5/5	3/5	5/5
F-measure	2/5	0/5	0/5	0/5	3/5	5/5	4/5	5/5	3/5	5/5
Sensitivity	2/5	5/5	1/5	0/5	3/5	4/5	2/5	2/5	3/5	4/5
Specificity	1/5	1/5	1/5	1/5	3/5	0/5	2/5	0/5	3/5	5/5

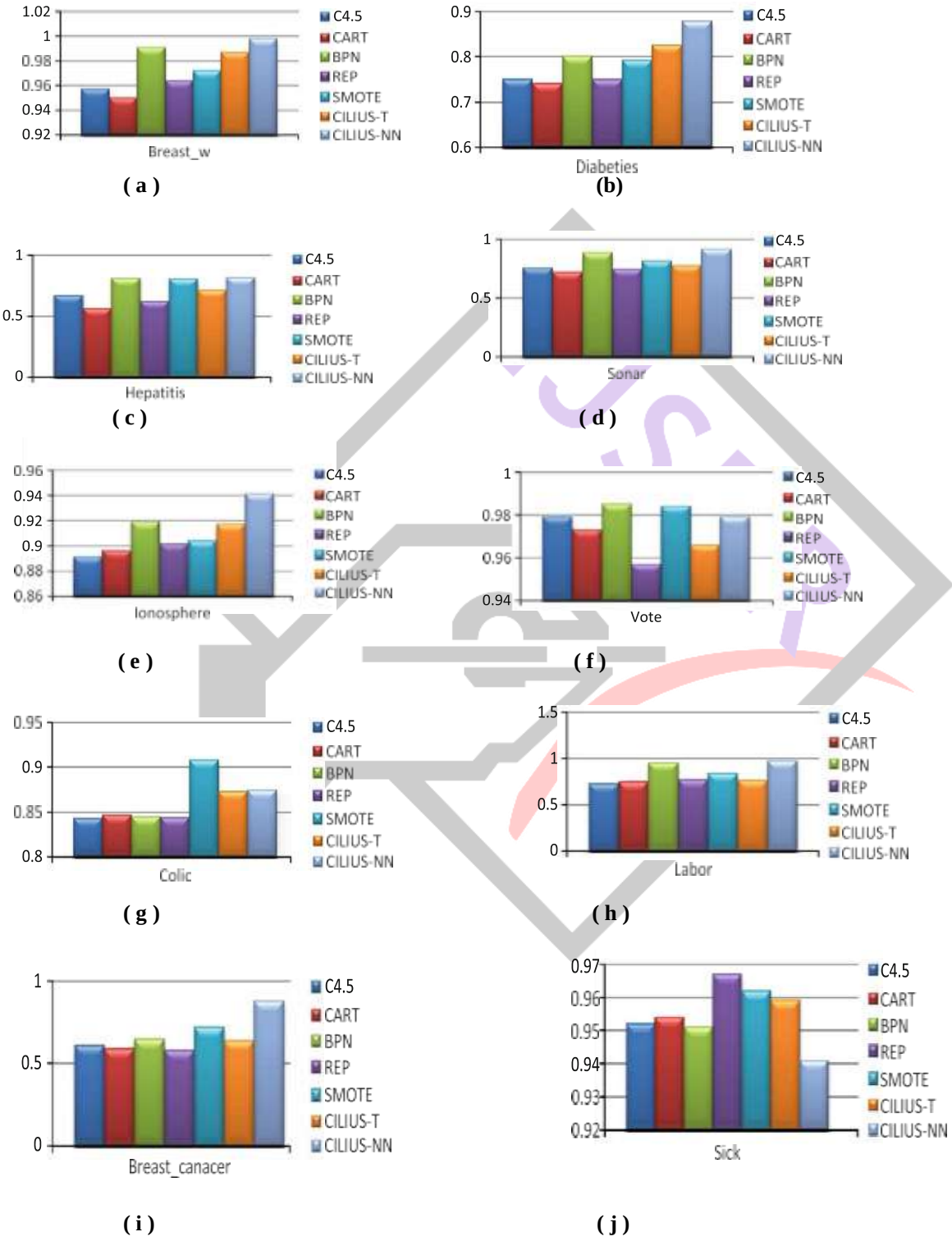
7. Results

We evaluated the performance of the proposed CILIUS 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. Experiments on the UCI datasets have two goals. First, we study the class imbalance properties of the datasets using proposed CILIUS learning algorithms. Second, we com- pare the classification performance of our proposed CILIUS algorithm with the traditional and class imbalance learning methods based on all datasets.

Following, we analyze the performance of the methods considering the entire original algorithms, without pre- processing, datasets for C4.5, CART, BPN and REP. we also analyze a pre-processing method SMOTE for performance evaluation of CILIUS variations. The complete table of results for all the algorithms used in this study is shown in Tables 3, 4, 5, 6, 7, 8 and 9, where the reader can observe the full test results, of performance of each approach with their associated standard deviation. The average result for each algorithm is given in the last row for each measure. We must emphasize the good results achieved by CILIUS-T and CILIUS-NN, as it obtains the highest value among all algorithms. In order to analyze these results, Fig. 2a–j shows the average AUC computed for all approaches, where we can observe that CILIUS-T and CILIUS-NN have obtained the best AUC value in the comparison and therefore it is clearly given the indication of its supremacy. Tables 8 and 9 presents the summary of experimental results conducted using non-parametric statistical tests with Wilcoxon test [50,51] for all the datasets. The table describes wins or tie of CILIUS-T and CILIUS-NN results of AUC, precision, F-measure, sensitivity and specificity for every dataset. The values *M/N* specify the *M* number of wins or tie on *N* algorithms compared.

Tables 8 and 9 presents the comparative results of pro- posed algorithms CILIUS-TT and CILIUS-NN against C4.5, CART, BPN, REP and SMOTE. The value in the table; exam- ple: "4/5" specifies that the proposed algorithm has registered four wins against compared five algorithms on that dataset for that specified measure. One can observe from the Tables 8 and 9 that our proposed algorithms have registered good number of wins against the compared algorithms on all the datasets.

Fig. 2 a–j Test results on AUC between the C4.5, CART, BPN, REP, SMOTE, CILIUS-T and CILIUS-NN for Breast_w, Diabetes, Hepatitis, Sonar, Ionosphere, Vote, Colic, Labor, Breast_cancer and Sick datasets



8. 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 two variations using C4.5 and Back propagation neural networks as its base learners. Experimental results show that CILIUS-T and CILIUS-NN have performed well in the case of multi class imbalance datasets. Furthermore, CILIUS is much less volatile than C4.5 and Back propagation neural networks original method. In our future work, we will apply CILIUS to more learning tasks, especially high dimensional feature learning tasks.

References

[1] Wu, J., Brubaker, S.C., Mullin, M.D., Rehg, J.M.: Fast asymmet- ric learning for cascade face detection. IEEE Trans. Pattern Anal. Mach. Intell. 30(3), 369–382 (2008)
[2] Chawla, N.V., Japkowicz, N., Kotcz, A. (eds.): Proceedings of ICML Workshop Learn. Imbalanced Data Sets (2003)
[3] Japkowicz, N. (ed.) Proceedings of AAAI Workshop Learn. Imbal- anced Data Sets (2000)
[4] Weiss, G.M.: Mining with rarity: a unifying framework. ACM SIGKDD Explor. Newslett. 6(1), 7–19 (2004)
[5] Chawla, N.V., Japkowicz, N., Kolcz, A. (eds.): Special issue learn- ing imbalanced datasets. SIGKDD Explor. Newslett. 6(1) (2004)
[6] Lu, W.-Z., Wang, D.: Ground-level ozone prediction by sup- port vector machine approach with a cost-sensitive classification scheme. Sci. Total. Environ. 395(2–3), 109–116 (2008)
[7] Huang, Y.-M., Hung, C.-M., Jiau, H.C.: Evaluation of neural net- works and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Anal. R. World Appl. 7(4), 720–747 (2006)
[8] Cieslak, D., Chawla, N., Striegel, A.: Combating imbalance in network intrusion datasets. In: IEEE International Confernce on Granular Computation, pp. 732–737 (2006)
[9] Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for med- ical decision making: the effects of imbalanced datasets on classi- fication performance. Neural Netw. 21(2–3), 427–436 (2008)
[10] Freitas, A., Costa-Pereira, A., Brazdil, P.: Cost-sensitive decision trees applied to medical data. In: Song, I., Eder, J., Nguyen, T. (eds.) Data Warehousing Knowl. Discov. (Lecture Notes Series in Computer Science) (2007)
[11] Kiliç, K., T ürksen, I.B.: Comparison of different strategies of uti- lizing fuzzy clustering in structure identification. Inf. Sci. 177(23), 5153–5162 (2007)
[12] Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslando- gan, Y.A., Stoecker, W.V., Moss, R.H.: A methodological approach to the classification of dermoscopy images. Comput. Med. Imag. Graph. 31(6), 362–373 (2007)
[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, vol. 4654, pp. 303–312, (2007).
[14] Batuwita, R., Palade, V.: FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Trans. Fuzzy Syst. 18(3), 558–571 (2010)
[15] Japkowicz, N., Stephen, S.: The class imbalance problem: a sys- tematic study. Intell. Data Anal. 6, 429–450 (2002)
[16] Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: Proceedings of the 14th International Conference on Machine Learning, pp. 179–186 (1997)
[17] Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A Study of the Behavior of Several Methods for Balancing Machine Learn- ing Training Data, SIGKDD Explorations, vol. 6, pp. 20–29 (2004)
[18] Cieslak, D., Chawla, N.: Learning decision trees for unbalanced data. In: Machine Learning and Knowledge Discovery in Data- bases. Springer, Berlin, pp. 241–256 (2008)
[19] Weiss, G.: Mining with rarity: a unifying framework. SIGKDD Explor. Newslett. 6(1), 7–19 (2004)
[20] Chawla, N., Bowyer, K., Kegelmeyer, P.: SMOTE: synthetic minor- ity over-sampling technique. J. Artif. Intell. Res. 16, 321–357 (2002)
[21] Zhang, J., Mani, I.: KNN approach to unbalanced data distribu- tions: a case study involving information extraction. In: Proceed- ings of the International Conference on Machine Learning, Work- shop: Learning Imbalanced Data Sets, Washington, DC, pp. 42–48 (2003)
[22] Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. ACM SIGKDD Explor. Newslett. 6(1), 40–49 (2004)
[23] Zou, S., Huang, Y., Wang, Y., Wang, J., Zhou, C.: SVM learning from imbalanced data by GA sampling for protein domain predic- tion. In: Proceedings of the 9th International Conference on Young Computer Scientists, Hunan, pp. 982–987 (2008)
[24] Wang, J., You, J., Li, Q., Xu, Y.: Extract minimum positive and maximum negative features for imbalanced binary classification. Pattern Recognit. 45, 1136–1145 (2012)
[25] Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39, 3446–3453 (2012)
[26] Garcì, S., Derrac, J., Triguero, I., Carmona, C.J., Herrera, F.: Evolutionary-based selection of generalized instances for imbal- anced classification. Knowl. Based Syst. 25, 3–12 (2012)
[27] Xiao, J., Xie, L., He, C., Jiang, X.: Dynamic classifier ensemble model for customer classification with imbalanced class distribu- tion. Expert Syst. Appl. 39, 3668–3675 (2012)
[28] López, V., Fernández, A., Moreno-Torres, J.G., Herrera, F.: Analy- sis of preprocessing vs. cost-sensitive learning for imbalanced clas- sification. Open problems on intrinsic data characteristics. Expert Syst. Appl. 39, 6585–6608 (2012)
[29] Yong, Y.: The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm. Energy Procedia 17, 164–170 (2012)

[30] Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A.: RUS- Boost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 40(1), 185 (2010)
[31] Garcia, V., Sanchez, J.S., Mollineda, R.A.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowl. Based Syst. 25, 13–21 (2012)
[32] Pérez-Godoy, M.D., Fernández, A., Rivera, A.J., del Jesus, M.J.: Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets. Pattern Recognit. Lett. 31, 2375–2388 (2010)
[33] Li, D.C., Liu, C.W., Hu, S.C.: A learning method for the class imbalance problem with medical data sets. Comput. Biol. Med. 40, 509–518 (2010)
[34] Che, E., Lin, Y., Xiong, H., Luo, Q., Ma, H.: Exploiting prob- abilistic topic models to improve text categorization under class imbalance. Inf. Process. Manag. 47, 202–214 (2011)
[35] Fernández, A., del Jesus, M.J., Herrera, F.: On the 2-tuples based genetic tuning performance for fuzzy rule based classification sys- tems in imbalanced data-sets. Inf. Sci. 180, 1268–1291 (2010)
[36] Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition. World Scientific (1996)
[37] Ishibuchi, H., Yamamoto, T., Nakashima, T.: Hybridization of fuzzy GBML approaches for pattern classification problems. IEEE Trans. Syst. Man Cybern. B 35(2), 359–365 (2005)
[38] Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Syst. Appl. 36, 4626–4636 (2009)
[39] Hsu, C.C., Wang, K.S., Chang, S.H.: Bayesian decision theory for support vector machines: imbalance measurement and feature opti- mization. Expert Syst. Appl. 38, 4698–4704 (2011)
[40] Fernández, A., del Jesus, M.J., Herrera, F.: On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets. Expert Syst. Appl. 36, 9805–9812 (2009)
[41] Malof, J.M., Mazurowski, M.A., Tourassi, G.D.: The effect of class imbalance on case selection for case-based classifiers: an empirical study in the context of medical decision support. Neural Netw. 25, 141–145 (2012)
[42] Fernández, A., del Jesus, M.J., Herrera, F.: On the 2-tuples based genetic tuning performance for fuzzy rule based classification sys- tems in imbalanced data-sets. Inf. Sci. 180, 1268–1291 (2010)
[43] Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition. World Scientific (1996)
[44] Ishibuchi, H., Yamamoto, T., Nakashima, T.: Hybridization of fuzzy GBML approaches for pattern classification problems. IEEE Trans. Syst. Man Cybern. B 35(2), 359–365 (2005)
[45] Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Syst. Appl. 36, 4626–4636 (2009)
[46] Hsu, C.C., Wang, K.S., Chang, S.H.: Bayesian decision theory for support vector machines: imbalance measurement and feature opti- mization. Expert Syst. Appl. 38, 4698–4704 (2011)
[47] Fernández, A., del Jesus, M.J., Herrera, F.: On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets. Expert Syst. Appl. 36, 9805–9812 (2009)
[48] Malof, J.M., Mazurowski, M.A., Tourassi, G.D.: The effect of class imbalance on case selection for case-based classifiers: an empirical study in the context of medical decision support. Neural Netw. 25, 141–145 (2012)
[49] Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD Thesis, Hamilton (1998)
[50] Quinlan, J.R.: C4.5: Programs for Machine Learning, 1st edn. Morgan Kaufmann Publishers, San Mateo (1993)
[51] Rumelhart, D.E., Hinton, Geoffrey E., Williams, Ronald J., Learn- ing representations by back-propagating errors. Nature 323(6088), 533–536 (1986). doi:10.1038/323533a0
[52] Asuncion, A., Newmann, D.: UCI Repository of Machine Learning Database (School of Information and Computer Science) Univer- sity of California, Irvine (2007, online). http://www.ics.uci.edu/
[53] ~mlearn/MLRepository.html
[54] Fernández, A., García, S., del Jesus, M.J., Herrera, F.: A study on the use of the fuzzy reasoning method based on the winning rule vs. voting procedure for classification with imbalanced data sets. In: Ninth International Work-conference on Artificial Neural Networks (IWANN07). Lecture Notes on Computer Science, vol. 4507, pp. 375–382. Springer, Berlin (2007)
[55] Fernández, A., García, S., del Jesus, M.J., Herrera, F.: A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets Syst. 159(18), 2378–2398 (2008)
[56] García, S., Fernández, A., Luengo, J., Herrera, F.: A study of sta- tistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Comput. (2009). doi:10.1007/s00500-008-0392-y
[57] Orriols-Puig, A., Bernadó-Mansilla, E.: Evolutionary rule-based systems for imbalanced datasets. Soft Comput. 13(3), 213–225 (2009)
[58] García, S., Molina, D., Lozano, M., Herrera, F.: A study on the use of non-parametric tests for analyzing the evolutionary algo- rithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. J Heurist. 15, 617–644 (2009). doi:10.1007/s10732-008-9080-4
[59] Luengo, J., García, S., Herrera, F.: A study on the use of statistical tests for experimentation with neural networks: analysis of para- metric test conditions and non-parametric tests. Expert Syst. Appl. 36(2009), 7798–7808 (2009)