A New Approach for Cluster Disjuncts using Naive Bayes

Syed Ziaur Rahman, G. Samuel Vara Prasad Raju

Abstract— **Data mining is the process of discovering hidden knowledge from the existing databases. In real-time applications, most often data sources are of imbalanced nature. 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 Naive Bayes Cluster Disjunct (NBCD) for learning from skewed training data. These algorithms provide a simpler and faster alternative by using naive bayes as base algorithm. We conducted experiments using fifteen UCI data sets from various application domains using five algorithms for comparison on six 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.**

Index Terms* — *Classification, class imbalance, Cluster Disjunct, NBCD.

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

II. Literature Review

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:

In [14] authors proposed a weighted online sequential extreme learning machine (WOS-ELM) algorithm for class imbalance learning (CIL). WOS-ELM is a general online learning method that alleviates the class imbalance problem in both chunk-by-chunk and one-by-one learning. One of the new features of WOS-ELM is that an appropriate weight setting for CIL is selected in a computationally efficient manner. In [15] authors proposes a methodology to find a (near-) optimal class distribution for class imbalance data sources. One more aim of the authors is to show that balancing the class distribution is not always the best solution when intelligent resampling methods are used, i.e. there is often a class distribution other than 50 % that improves the results. They presented a methodology to find a (near-) optimal class distribution. In [16] authors presented a new approach for dealing with class-imbalanced datasets based on a new boosting method for the construction of ensembles of classifiers. The approach is based on using the distribution of the weights given by a given boosting algorithm for obtaining a supervised projection. Then, the supervised projection is used to train the next classifier using a uniform distribution of the training instances.

Manuscript received July 10, 2014; revised August 06, 2014.

Syed Ziaur Rahman is currently working as a faculty with Ministry of Higher Education and also a part time PhD Scholar with Andhra University, Vishakhapatnam, Andhra Pradesh, India. Email: sdzrahman@gmail.com Phone: 0091 9885615794

G. Samuel Vara Prasad Raju is currently working as an Associate Professor with Andhra University, Vishakhapatnam, Andhra Pradesh, India. Email: gsvpraju2012@gmail.com

In [17] authors have proposed the use of three approaches to surrounding neighborhood with the aim of generating artificial minority instances, but taking into account both the proximity and the spatial distribution of the examples. The topics discussed in Section 2 provide the foundation for most of the current research activities on imbalanced learning.

III. Proposed NBCD Framework

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 oversampling tries to generate synthetic minority instances. Before performing oversampling on the minority subset, the main cluster disjuncts has to be identified and the borderline and noise instances around the cluster disjuncts are to be removed. The number of instances eliminated will belong to the 'k' cluster disjuncts selected by visualization technique. The remaining cluster disjunct instances have to be oversampled by using hybrid synthetic oversampling technique. Here, the above said routine is employed on every cluster disjunct, which removes examples suffering from missing values at first and then removes borderline examples and examples of outlier category.

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

A. *Preparation of the Majority and Minority subsets*

The datasets is partitioned into majority and minority subsets. As we are concentrating over sampling, we will take minority data subset for further visualization analysis to identify cluster disjuncts.

B. *Initial phase of removing noisy and cluster disjunct borderline instances*

Minority subset can be further analyzed to find the noisy or borderline 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 choose the noisy instances relating to that cluster disjunct 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 cluster disjunct. This process can be applied on all the cluster disjuncts identified for each dataset.

C. *Applying oversampling on cluster disjunct*

The oversampling of the instances can be done on the improved cluster disjuncts produced in the earlier phase. The oversampling can be done as follows:

Apply resampling supervised filter on the cluster disjunct for generating synthetic instances. The synthetic minority instances generated can have a percentage of instances which can be replica of the pure instances and reaming percentage of instances are of the hybrid quality of synthetic instances generated by combing two or more instances from the pure minority sunset. Perform oversampling on cluster disjunct can help so as to form strong, efficient and more valuable rules for proper knowledge discovery.

D. *Forming the strong dataset*

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

IV. Experimental framework

In this section we first describe the collection of imbalanced data sets selected for our study and corresponding parameters for experimental setup.

In this study our proposed algorithm is applied to fifteen binary data sets from the UCI repository [18] with different imbalance ratio (IR). Table 1 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.

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, 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, AUC, Precision, F-measure, as performance evaluation measures.

TABLE I
Summary of benchmark imbalanced datasets

S.no	Datasets	# Ex.	# Atts.	Class (_,+)	IR
1.	Breast	268	9	(recurrence; no-recurrence)	2.37
2.	Breast_w	699	9	(benign; malignant)	1.90
3.	Colic	368	22	(yes; no)	1.71
4.	Credit-g	1000	21	(good; bad)	2.33
5.	Diabetes	768	8	(tested-potv; tested-negtv)	1.87
6.	Heart-c	303	14	(<50,>50_1)	1.19
7.	Heart-h	294	14	(<50,>50_1)	1.77
8.	Heart-stat	270	14	(absent, present)	1.25
9.	Hepatitis	155	19	(die; live)	3.85
10.	Ionosphere	351	34	(b;g)	1.79
11.	Kr-vs-kp	3196	37	(won; nowin)	1.09
12.	Labor	56	16	(bad ; good)	1.85
13.	Mushroom	8124	23	(e ; p)	1.08
14.	Sick	3772	29	(negative ; sick)	15.32
15.	Sonar	208	60	(rock ; mine)	1.15

. In order to estimate these different measure 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.

.

V. Results

We compared proposed method with the SVM, C4.5 [19], NN [20], FT and SMOTE [21] 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 analyze the performance of the method considering the entire original algorithms, without pre-processing, data sets for SVM, C4.5, NN and FT. we also analyze a pre-processing method SMOTE for performance evaluation of proposed algorithm.

The complete table of results for all the algorithms used in this study is shown in Table II, 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 our proposed algorithm, as it obtains the highest value among all algorithms

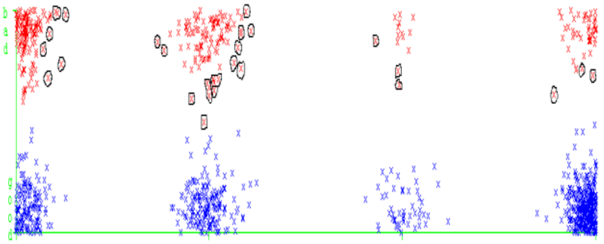

Fig. 1. Before applying NBCD on credit-g data set

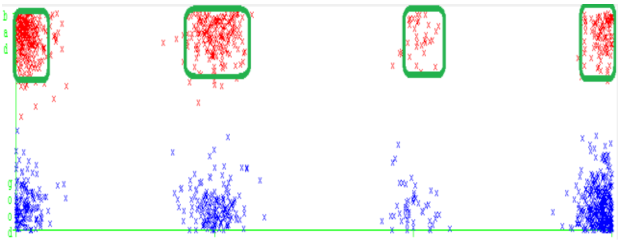

Fig. 2. After applying NBCD on credit-g data set

Table II reports the results of accuracy, AUC, precision, F-measure, TP Rate and TN Rate respectively for Breast, Breast_w, Colic, Credit-g, Diabetes, Heart-c, Heart-h, Heart-stat, Hepatitis, Ionosphere, Kv-rs-kp, Labor, Mushroom , Sick and Sonar datasets. The bullet '●' indicates a win of proposed method on SVM, C4.5, NN, FT and SMOTE and a circle '○' indicates a loss of our proposed method on above said algorithms. The results in the tables show that our proposed method 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 our proposed method. A two-tailed corrected resampled paired t-test 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.

Table III reports the comparison of our proposed approach with a recent published algorithm CILIUS [22] and our proposed algorithm has performed well. Finally, we can say that our proposed method is one of the best alternatives to handle class imbalance problems effectively.

This experimental study supports the conclusion that the an learning algorithm should know all the minority small sub concepts improved CIL behavior when dealing with imbalanced data-sets, as it has helped the proposed method to be the best performing algorithms when compared with four classical and well-known algorithms: SVM, C4.5, NN, FT and a well-established pre-processing technique SMOTE.

VI. 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 a new hybrid frame work dubbed as Naive Bayes Cluster Disjunct (NBCD) for learning from skewed training data.

These algorithms provide a simpler and faster alternative by using naive bayes as base algorithm. Experimental results show that NBCD has performed well in the case of multi class imbalance datasets. Furthermore, NBCD is much less volatile than C4.5. In our future work, we will apply NBCD to more learning tasks, especially high dimensional feature learning tasks. Another variation of our approach in future work is to analyze the influence of same base classifier and different base classifier effect on the quality of synthetic minority instances generated.

TABLE II
SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE FOR PROPOSED ALGORITHM ON ALL THE DATASETS

Datasets	SMOTE	C4.5	NN	FT	SVM	NBCD
Accuracy						
Breast	69.83±7.77●	69.52±7.50●	74.28±6.05○	68.58±7.52●	67.21±7.28●	73.356±6.603
Breast_w	96.16±2.06●	96.75±2.01●	95.01±2.73●	95.45±2.52●	96.75±2.00●	97.971±1.503
Colic	88.53±4.10●	82.66±5.41●	85.16±5.91●	79.11± 6.51●	79.78±6.57●	90.641±4.640
Credit-g	76.50±3.38	75.09±3.42●	71.25±3.17●	71.88±3.68●	68.91±4.46●	76.844±4.494
Diabetes	76.08±4.04●	76.80±4.54●	74.49±5.27●	70.62± 4.67●	76.55±4.67●	79.333±4.137
Heart-c	82.99±4.98	83.86±6.21	76.94±6.59●	76.06±6.84●	81.02±7.25●	83.052±6.371
Heart-h	85.65±5.46	82.74±6.44●	80.22±7.95●	78.33±7.54●	81.81±6.20●	85.178±5.143
Heart-stat	83.89±5.05○	83.89±6.24○	78.15±7.42●	76.15±8.46●	82.07±6.88○	81.872±7.342
Hepatitis	78.35±9.09●	85.77±9.04●	79.22±9.57●	81.40±8.55●	81.90±8.38●	89.529±8.001
Ionosphere	90.28±4.73●	88.07±5.32●	89.74±4.38●	87.10±5.12●	90.26±4.97●	94.411±3.590
Kv-rs-kp	99.66±0.27○	95.79±1.34●	99.44±0.37○	90.61±1.65●	99.02±0.54○	98.103±1.636
Labor	80.27±11.94●	92.97±9.75●	78.60±16.58●	84.30±16.24●	92.40±11.07●	95.905±7.259
Mushroom	100.0±0.00	100.0±0.00	100.0±0.00	100.0± 0.00	100.0±0.00	100.0±0.00
Sick	97.61±0.68●	93.87±0.13●	98.72±0.55	96.10±0.92●	99.26±0.04○	98.379±0.691
Sonar	82.42±7.25●	76.60±8.27●	73.61±9.34●	86.17±8.45●	75.46±9.92●	86.107±8.187
AUC						
Breast	0.717±0.084●	0.584±0.086●	0.606±0.087●	0.604±0.082●	0.586±0.102●	0.799±0.074
Breast_w	0.967±0.025●	0.966±0.023●	0.957±0.034●	0.949±0.030○	0.977±0.017●	0.991±0.009
Colic	0.908±0.040●	0.810±0.060●	0.843±0.070●	0.777±0.072●	0.802±0.073●	0.958±0.029
Credit-g	0.778±0.041●	0.670±0.043●	0.647±0.062●	0.655±0.044●	0.650±0.075●	0.847±0.043
Diabetes	0.791±0.041●	0.713±0.055●	0.751±0.070●	0.668±0.051●	0.793±0.072●	0.849±0.040
Heart-c	0.830±0.077●	0.834±0.064●	0.769±0.082●	0.757±0.069●	0.843±0.084●	0.913±0.052
Heart-h	0.904±0.054●	0.797±0.074●	0.775±0.089●	0.763±0.082●	0.852±0.078●	0.923±0.043
Heart-stat	0.832±0.062●	0.834±0.064●	0.786±0.094●	0.760±0.085●	0.864±0.075●	0.870±0.068
Hepatitis	0.798±0.112●	0.768±0.144●	0.668±0.184●	0.678±0.139●	0.757±0.195●	0.952±0.056
Ionosphere	0.904±0.053●	0.845±0.069●	0.891±0.060●	0.831±0.067●	0.900±0.060●	0.961±0.032
Kr-vs-kp	0.999±0.001○	0.9580.014●	0.998±0.003○	0.906±0.017●	0.996±0.005○	0.995±0.004
Labor	0.833±0.127●	0.917±0.122●	0.726±0.224●	0.844±0.162●	0.971±0.075○	0.995±0.024
Mushroom	1.000±0.000	1.000±0.00	1.000±0.000	1.000±0.00	1.000±0.000	1.000±0.000
Sick	0.962±0.025●	0.501±0.005●	0.952±0.040●	0.795±0.053●	0.990±0.014●	0.979±0.019
Sonar	0.814±0.090●	0.764±0.083●	0.753±0.113●	0.859±0.086●	0.771±0.103●	0.924±0.063
Precision						
Breast	0.710±0.075●	0.747±0.048●	0.753±0.042○	0.762±0.051●	0.745±0.051●	0.770±0.062
Breast_w	0.974±0.025●	0.979±0.021●	0.965±0.026●	0.964±0.026●	0.988±0.019●	0.996±0.011
Colic	0.853±0.057●	0.857±0.053●	0.851±0.055●	0.839±0.062●	0.845±0.060●	0.925±0.058
Credit-g	0.768±0.034●	0.793±0.026●	0.767±0.025●	0.791±0.027●	0.776±0.033●	0.805±0.052
Diabetes	0.781±0.064●	0.782±0.038●	0.797±0.045●	0.764±0.036●	0.793±0.037●	0.826±0.054
Heart-c	0.779±0.082●	0.832±0.070	0.783±0.076●	0.776±0.068●	0.825±0.080●	0.831±0.084
Heart-h	0.878±0.076●	0.841±0.058●	0.824±0.071●	0.830±0.063●	0.849±0.058●	0.896±0.070
Heart-stat	0.791±0.081●	0.846±0.070○	0.799±0.051●	0.796±0.085●	0.833±0.078○	0.828±0.084
Hepatitis	0.709±0.165●	0.710±0.278○	0.510±0.371●	0.546±0.333●	0.604±0.271●	0.791±0.151
Ionosphere	0.934±0.049●	0.938±0.072●	0.895±0.084●	0.938±0.073●	0.906±0.080●	0.944±0.051
Kr-vs-kp	0.996±0.005○	0.963±0.019●	0.994±0.006○	0.905±0.021●	0.991±0.008○	0.978±0.023
Labor	0.871±0.151●	0.932±0.181●	0.696±0.359●	0.802±0.250●	0.915±0.197●	0.938±0.122
Mushroom	1.000±0.000	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.000	1.000±0.000
Sick	0.983±0.007●	0.939±0.001●	0.992±0.005○	0.975±0.007●	0.997±0.003○	0.990±0.005
Sonar	0.863±0.068○	0.767±0.107●	0.728±0.121●	0.883±0.100○	0.764±0.119●	0.858±0.092
F-measure						
Breast	0.730±0.076●	0.797±0.054○	0.838±0.040○	0.776±0.057●	0.781±0.059●	0.782±0.056
Breast_w	0.960±0.022●	0.975±0.015●	0.962±0.021●	0.975±0.016●	0.965±0.019●	0.980±0.015
Colic	0.880±0.042●	0.863±0.044●	0.888±0.044●	0.838±0.054●	0.833±0.055●	0.908±0.045
Credit-g	0.787±0.034○	0.830±0.024○	0.805±0.022○	0.779±0.034●	0.802±0.027○	0.784±0.041
Diabetes	0.741±0.046●	0.834±0.033○	0.806±0.044○	0.827±0.038○	0.778±0.037●	0.786±0.044
Heart-c	0.772±0.070●	0.858±0.053○	0.792±0.059●	0.827±0.069	0.782±0.064●	0.827±0.065
Heart-h	0.841±0.061○	0.870±0.049○	0.851±0.061○	0.859±0.052○	0.830±0.063●	0.850±0.054
Heart-stat	0.789±0.072●	0.858±0.055○	0.806±0.069●	0.791±0.072●	0.781±0.083●	0.819±0.077
Hepatitis	0.677±0.138●	0.630±0.235●	0.409±0.272●	0.557±0.207●	0.469±0.265●	0.830±0.129
Ionosphere	0.905±0.048●	0.807±0.095●	0.850±0.066●	0.855±0.079●	0.787±0.098●	0.942±0.037
Kv-rs-kp	0.995±0.004○	0.960±0.013●	0.995±0.004○	0.991±0.005○	0.911±0.016●	0.981±0.016
Labor	0.793±0.132●	0.881±0.189●	0.636±0.312●	0.879±0.195●	0.794±0.211●	0.954±0.082
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.987±0.004●	0.968±0.001●	0.993±0.003○	0.996±0.003○	0.979±0.005●	0.991±0.004
Sonar	0.861±0.061●	0.743±0.095●	0.716±0.105●	0.753±0.102●	0.844±0.099●	0.866±0.080

			TP Rate			
Breast	0.763±0.117●	0.860±0.085○	0.947±0.060○	0.815±0.095●	0.806±0.091○	0.800±0.085
Breast_w	0.947±0.035●	0.972±0.025○	0.959±0.033○	0.962±0.029●	0.967±0.025○	0.965±0.026
Colic	0.913±0.058○	0.873±0.065○	0.931±0.053○	0.835±0.077●	0.832±0.075○	0.896±0.063
Credit-g	0.783±0.052○	0.872±0.039○	0.847±0.036○	0.810±0.058○	0.815±0.041○	0.767±0.051
Diabetes	0.868±0.065○	0.894±0.046○	0.821±0.073○	0.712±0.089●	0.795±0.054○	0.753±0.061
Heart-c	0.837±0.100○	0.889±0.068○	0.808±0.085●	0.777±0.110●	0.795±0.095●	0.831±0.092
Heart-h	0.876±0.089○	0.906±0.072○	0.885±0.081○	0.815±0.084●	0.835±0.093○	0.816±0.088
Heart-stat	0.857±0.090○	0.875±0.079○	0.824±0.104○	0.803±0.110●	0.775±0.113●	0.817±0.102
Hepatitis	0.573±0.248●	0.617±0.270●	0.374±0.256●	0.681±0.188●	0.448±0.273●	0.892±0.149
Ionosphere	0.820±0.114●	0.718±0.131●	0.821±0.107●	0.881±0.071●	0.689±0.131●	0.943±0.053
Kv-rs-kp	0.990±0.007○	0.956±0.016●	0.995±0.005○	0.995±0.006○	0.916±0.021●	0.985±0.012
Labor	0.885±0.234●	0.875±0.240●	0.640±0.349●	0.765±0.194●	0.845±0.243●	0.983±0.073
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.00	1.000±0.000
Sick	0.995±0.004○	1.000±0.00●	0.995±0.004○	0.990±0.005●	0.984±0.006●	0.992±0.005
Sonar	0.757±0.136●	0.737±0.135●	0.721±0.140●	0.865±0.090●	0.820±0.131●	0.883±0.105
			TN Rate			
Breast	0.335±0.166●	0.307±0.148●	0.260±0.141●	0.403±0.144●	0.622±0.137●	0.634±0.128
Breast_w	0.977±0.037●	0.960±0.042●	0.932±0.052○	0.930±0.052●	0.975±0.024●	0.996±0.012
Colic	0.734±0.118●	0.746±0.106●	0.717±0.119●	0.721±0.123●	0.862±0.063●	0.918±0.069
Credit-g	0.469±0.098●	0.467±0.084●	0.398±0.085●	0.495±0.077●	0.713±0.056●	0.771±0.073
Diabetes	0.574±0.095●	0.532±0.100●	0.603±0.111●	0.540±0.086●	0.807±0.077●	0.834±0.063
Heart-c	0.779±0.117●	0.778±0.109●	0.723±0.119●	0.720±0.106●	0.861±0.068○	0.830±0.097
Heart-h	0.714±0.131●	0.688±0.133●	0.655±0.158●	0.690±0.139●	0.894±0.074○	0.891±0.085
Heart-stat	0.775±0.123●	0.793±0.109●	0.728±0.131●	0.744±0.124●	0.862±0.064○	0.820±0.098
Hepatitis	0.882±0.092●	0.920±0.086○	0.900±0.097○	0.909±0.086○	0.837±0.109●	0.896±0.090
Ionosphere	0.949±0.046○	0.972±0.033○	0.940±0.055●	0.973±0.032○	0.928±0.057●	0.945±0.054
Kv-rs-kp	0.990±0.009○	0.960±0.022●	0.993±0.007●	0.895±0.026●	0.998±0.003○	0.977±0.025
Labor	0.945±0.131●	0.959±0.110○	0.865±0.197●	0.843±0.210●	0.847±0.187●	0.946±0.106
Mushroom	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.974±0.026○	0.001±0.010●	0.875±0.071●	0.606±0.106●	0.872±0.053●	0.919±0.045
Sonar	0.752±0.148●	0.791±0.118●	0.749±0.134●	0.898±0.094○	0.752±0.113●	0.839±0.120

TABLE III
SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE FOR PROPOSED ALGORITHM ON ALL THE DATASETS

Datasets	CILIUS [22]	NBCD
	AUC	
Breast	0.637 ± 0.110	**0.799±0.074**
Breast_w	0.987 ± 0.016	**0.991±0.009**
Colic	0.873 ± 0.082	**0.958±0.029**
Diabetes	0.826 ± 0.056	**0.849±0.040**
Hepatitis	0.714 ± 0.166	**0.952±0.056**
Ionosphere	0.917 ± 0.048	**0.961±0.032**
Labor	0.765 ± 0.217	**0.995±0.024**
Sick	0.950 ± 0.035	**0.979±0.019**
Sonar	0.774 ± 0.114	**0.924±0.063**
	Precision	
Breast	0.736 ± 0.050	**0.770±0.062**
Breast_w	0.986 ± 0.020	**0.996±0.011**
Colic	0.787 ± 0.090	**0.925±0.058**
Diabetes	0.810 ± 0.048	**0.826±0.054**
Hepatitis	0.698 ± 0.305	**0.791±0.151**
Ionosphere	0.922 ± 0.071	**0.944±0.051**
Labor	0.754 ± 0.337	**0.938±0.122**
Sick	0.990 ± 0.006	0.990±0.005
Sonar	0.759 ± 0.112	**0.858±0.092**
	F-measure	
Breast	**0.812 ± 0.046**	0.782±0.056
Breast_w	**0.984 ± 0.014**	0.980±0.015
Colic	0.827 ± 0.073	**0.908±0.045**
Diabetes	**0.836 ± 0.040**	0.786±0.044
Hepatitis	0.556 ± 0.238	**0.830±0.129**
Ionosphere	0.881 ± 0.065	**0.942±0.037**
Labor	0.697 ± 0.307	**0.954±0.082**
Sick	0.991 ± 0.004	0.991±0.004
Sonar	0.752 ± 0.103	**0.866±0.080**
	TP Rate	
Breast	0.325 ± 0.156	**0.800±0.085**
Breast_w	**0.978 ± 0.030**	0.965±0.026
Colic	0.765 ± 0.122	**0.896±0.063**
Diabetes	0.696 ± 0.096	**0.753±0.061**
Hepatitis	**0.920 ± 0.092**	0.892±0.149
Ionosphere	**0.948 ± 0.052**	0.943±0.053
Labor	0.865 ± 0.207	**0.983±0.073**
Sick	0.903 ± 0.060	**0.992±0.005**
Sonar	0.743 ± 0.138	**0.883±0.105**

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, pp. 369–382, Mar. 2008.

[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, pp. 7–19, Jun. 2004.

[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, pp. 109–116, 2008.

[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, pp. 720–747, 2006.

[8] D. Cieslak, N. Chawla, and A. Striegel, “Combating imbalance in network intrusion datasets,” in 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, pp. 427–436, 2008.

[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, pp. 5153–5162, 2007.

[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, pp. 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, 2007, vol. 4654, pp. 303–312.

[14] Bilal Mirza, Zhiping Lin, and Kar-Ann Toh.Weighted online sequential extreme learn- ing machine for class imbalance learning. Neural Processing Letters, page (To Appear), 2013.

[15] Iñaki Albisua, Olatz Arbelaitz · Ibai Gurrutxaga, Aritz Lasarguren · Javier Muguerza · Jesús M. Pérez. “The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets”, Prog Artif Intell (2013) 2:45–63. DOI 10.1007/s13748-012-0034-6

[16] Nicolás García-Pedrajas · César García-Osorio. “Boosting for class-imbalanced datasets using genetically evolved supervised non-linear projections” Prog Artif Intell (2013) 2:29–44 DOI 10.1007/s13748-012-0028-4.

[17] V. García · J. S. Sánchez · R. Martín-Félez · R. A. Mollineda.” Surrounding neighborhood-based SMOTE for learning from imbalanced data sets”, Prog Artif Intell (2012) 1:347–362. DOI 10.1007/s13748-012-0027-5

[18] A. Asuncion D. Newman. (2007). 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

[19] J. R. Quinlan, C4.5: Programs for Machine Learning, 1st ed. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[20] Rumelhart, David E.; Hinton, Geoffrey E., Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". Nature 323 (6088): 533–536. doi:10.1038/323533a0.

[21] N. Chawla, K. Bowyer, and P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[22] Satuluri Naganjaneyulu · Mrithyumjaya Rao Kuppa. “A novel framework for class imbalance learning using intelligent under-sampling”, Prog Artif Intell (2013) 2:73–84. DOI 10.1007/s13748-012-0038-2