International Journal of Database Theory and Application
Vol.10, No.1 (2017), pp.47-56
http://dx.doi.org/10.14257/ijdta.2017.10.1.05

A Novel Random Forest Approach Using Specific under Sampling Strategy

L. Surya Prasanthi[1], R. Kiran Kumar[2] and Kudipudi Srinivas[3]

[1]Research Scholar, Department of Computer Science, Krishna University, Machilipatnam, India
[2]Department of Computer Science, Krishna University, Machilipatnam, India
[3]Department of Computer Science & Engineering, V.R. Siddartha Engineering College, Vijayawada, India
E-mail:prasanthi.latike@gmail.com, kirankreddi @gmail.com, vrdrks@gmail.com

Abstract

In Data Mining the knowledge is discovered from the existing real world data sets. In real time scenario, the category of datasets varies dynamically. One of the emerging categories of dataset is class imbalance data. In Class Imbalance data, the percentages of instances in one class are far greater than the other class. The traditional data mining algorithms are well applicable for knowledge discovery from balance datasets. Efficient knowledge discovery is hampered in the case of class imbalance datasets. In this paper, we propose a novel approach dubbed as Under Sampling using Random Forest (USRF) for efficient knowledge discovery from imbalance datasets. The proposed USRF approach is verified on the 11 benchmark datasets from UCI repository. The experimental observations show that an improved accuracy and AUC is achieved with the proposed USRF approach with a good reduction in RMS error.

Keywords*: Data Mining, Knowledge Discovery, Classification, Decision Tree, Imbalance Data, Random Forest, USRF*

1. Introduction

Data mining techniques can be broadly classified into classification and clustering. Classification is the process of classifying the labeled data into predefined classes. A general issue encountered in data mining is dealing with imbalance datasets, in which one class is predominantly outnumbers the other class. This issue results in high accuracy for the instances of majority class i.e. instances belonging to the predominant class and less accuracy for the instances of minority class. Therefore when dealing with class imbalance datasets a specific strategy has to be implemented for efficient knowledge discovery from the datasets. There are different type of approaches exists in the literature to handle the problem of class imbalance nature, to name a few are oversampling, under sampling, subset approaches, cost sensitive learning, algorithm level implementations and hybrid techniques which combine more than one approaches.

In oversampling, the instances in the minority subset are oversampled by following different strategies. In under sampling, the instances in the majority subset are reduced by several techniques. In subset approaches, the dataset is split into different subsets to reduce the imbalance nature. In cost sensitive learning, the instances are assigned with cost values and the reshuffling of the dataset is performed by considering the cost values. In algorithmic level approaches, the base algorithm applied to the class imbalance data is modified to suit with the imbalance data

learning. In hybrid level implementation, more than one above said approaches are applied to solve the problem of class imbalance learning. The existing approaches suffer

ISSN: 2005-4270 IJDTA

from the one or more of the drawbacks; either they performed the excessive oversampling, or/and they performed the excessive under sampling *etc*. We addressed the above issue by following a specific strategy for efficient under sampling using nearest neighbor technique. The results of experimental simulation show a good improvement against the benchmark traditional methods. To overall contributions of our work are as follows,

i. We presented the framework which shows how to pickup only a few instances for performing specific under sampling, and justify this selection process both theoretically and empirically.
ii. The proposed approach will work as a prototype for elaborating experimental analysis; due to the open availability of datasets, compared algorithms and evaluation measures *etc*.
iii. Finally, our proposed USRF approach outperform almost all the compared benchmark algorithms in terms of accuracy, AUC and root mean square error.

The reminder of the manuscript is organized as follows. Section 2 presets the related work connecting to the class imbalance learning. Section 3 initially presents the formal description of the framework and in the later section the algorithmic approach is also presented. Section 4 presents the experimental methodology and datasets. Section 5 elaborates the results of the proposed approach with the benchmark algorithms. Finally, is Section 6 conclusion is presented.

2. Class Imbalance Learning Approaches

This section, presented the summarized view of the recent proposals in the domain of class imbalance learning.

In [1] Iain Brown *et al.*, have conducted several experiments on credit scoring imbalance datasets and they shown that random forest is one of the best performing algorithm on the imbalance credit scoring datasets. In [2] Ana C. Lorena *et al.*, have applied machine learning algorithms especially random forest classifier for modeling species potential distribution.

Ehsan Molaei *et al.* [3] have developed a safe distributed algorithm which is using improved secure sum algorithm and performed on classic ID3.In [4] Yong Hu *et al.*, have investigated on stock trading techniques using the combined approaches of trend discovery and extended classifier system. In [5] Victoria López *et al.*, have proposed imbalance domain learning technique which uses iterative instance adjustment approach for efficient knowledge discovery. Sandeep Kumar *et al.* [6] have proposed an improved approach using ID3 as the base algorithm with Havrda and Charvatentrophy for building decision tree. Sagar Manohar *et al.*[7] have presented a classification approach for predicting future events.

In [8] Nele Verbiest *et al.*, have propose two prototype selection techniques both based on fuzzy rough set theory which removes noisy instances from the imbalanced dataset and generated synthetic instances. The above descried approaches are analyzed for discovering shortcomings and a novel algorithm know as under sampled random forest is proposed.

3. Framework of Under Sampled Radom Forest

This section presents the detail architecture of the proposed USRF approach which consists of four major modules. The detailed working principles of the USRF approach are explained below in the sub-sections.

In the initial phase of our proposed USRF the dataset is split into two subsets known as majority and minority subsets. The majority subset is the class of instances, which are more in percentage than the other class. The minority subset is the class of instances which are very less when compared to the other class in the dataset. Since the proposed approach is a under sampling approach. We considered the majority subset for more elaborate analysis for reduction of instances.

The instances in the majority subset are reduced by following the below mentioned techniques; one of the technique is to eliminate the noise instances, the other technique is to find the outliers and the final technique is to find the range of weak instances for removal. The noisy and outlier instances can be easily identified by analyzing the intrinsic properties of the instances. The range of weak instances can be identified by first identifying the weak features in the majority subset. The correlation based feature selection [9] technique selects the important features by following the inter correlation between feature - feature and the inter correlation between feature and class. The features which have very less correlation are identified for elimination. The range of instances which belong to these weak features are identified for elimination from the majority subset. The number of features and instances eliminate by the correlation based feature selection technique will vary from dataset to dataset depending upon the unique properties of the dataset.

The proposed USRF algorithm is summarized as below.

Algorithm: Under Sampled Radom Forest (USRF)

Algorithm: New Predictive Model

Input: D – Data Partition, A – Attribute List, GR – Gain Ratio

Output: A Decision Tree

Procedure:

Processing Phase:

Step 1. *Take the class imbalance data and divide it into majority and minority sub sets. Let the minority subset be P€ pi* (*i* = 1,2,..., *pnum*) *and majority subset be N € ni*(*i* = 1,2,..., *nnum*).

Let us consider
m' = the number of majority nearest neighbors
T= the whole training set
m= the number of nearest neighbors

Step 2. *Find mostly misclassified instances pi*
pi = m'; where m' ($0 \leq m' \leq m$)
if m/ $2 \leq m' < m$ *then pi is a mostly misclassified instance. Then remove the instances m' from the minority set.*

Let us consider
m' = the number of minority nearest neighbors
Step 3. *Find mostly misclassified instances ni*
ni = m'; where m' ($0 \leq m' \leq m$)
if m/ $2 \leq m' < m$ *then ni is a mostly misclassified instance. Then remove the instances m' from the majority set.*

Let us consider
m' = the number of majority nearest neighbors
Step 4. *Find noisy instances pi'*

pi' = m'; where m' (0 ≤ m'≤ m)
If m'= m, i.e. all the m nearest neighbors of pi are majority examples, pi' is considered to be noise or outliers or missing values and are to be removed.

Let us consider
m' = the number of minority nearest neighbors
***Step 5.** Find noisy instances ni'*
ni' = m'; where m' (0 ≤ m'≤ m)
If m'= m, i.e. all the m nearest neighbors of pi are minority examples, ni' is considered to be noise or outliers or missing values and are to be removed.

***Step 6.** In this step, we generate s × dnum synthetic minority examples from the minority sub set, where s is an integer between 1 and k . One percentage of synthetic examples generated is replica of minority examples and other are the hybrid of minority examples.*

Building Predictive Model:

```
Create a node N
  If samples in N are of same class, C then
    return N as a leaf node and mark class C;
      If A is empty then
return N as a leaf node and mark with majority class;
else
          apply Radom Forest
endif
endif
Return N
```

In the concluding phase of the algorithm, the subset in which irrelevant instances are removed is merged with the minority subset to form the strong dataset, which is further applied to the base algorithm for experimental simulation. In this context random forest [10] is used as the base algorithm for experimental simulation and results generation.

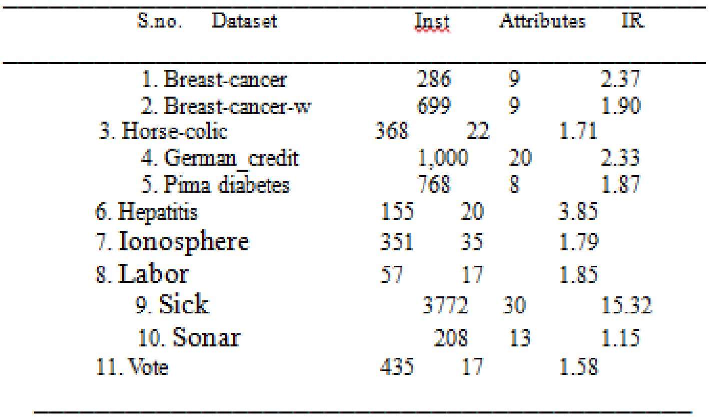
4. Methodology and Datasets

The methodology used for validation of generated experimental results is 10 fold cross validation. The 10 fold cross validation for 10 runs is considered as a decent validation set up in most of the benchmark empirical results simulation in the field of classification. Since, in the 10 fold cross validation the mean of 10 runs of each and every measure is considered, the precision of the results can be agreed on any terms. The proposed approach is compared with benchmark set of algorithms C4.5 [12], Reduced Error Pruning (REP) Tree [12], Classification and Regression Trees (CART) [13] and NB Tree [14]. The experiments are implemented within the weka [11] environment on windows 7, i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM.

Datasets used in Decision tree Learning

The datasets for the experiments are downloaded from the UCI [15] machine learning repository, which are described in the Table 1.

Table 1. UCI Datasets and their Properties



S.no. Dataset	Inst	Attributes	IR
1. Breast-cancer	286	9	2.37
2. Breast-cancer-w	699	9	1.90
3. Horse-colic	368	22	1.71
4. German_credit	1,000	20	2.33
5. Pima diabetes	768	8	1.87
6. Hepatitis	155	20	3.85
7. Ionosphere	351	35	1.79
8. Labor	57	17	1.85
9. Sick	3772	30	15.32
10. Sonar	208	13	1.15
11. Vote	435	17	1.58

The set of eleven UCI datasets: Breast-cancer, Breast-cancer-w, Horse-colic, German_credit, Pima diabetes, Hepatitis, Ionosphere, Labor, Sick, Sonar and Vote are used for experimental simulation. The Imbalance Ratio (IR) shown on the last column of the Table 1 gives the value of imbalance ratio. The value of IR can be calculated for the dataset by dividing the number of instances in the majority subset with number of instances in the minority subset.

The accuracy is the percentage of instances correctly classified by a classifier. The accuracy can also be defied in the terms of True positive (TP): Actual positive instances which are classified as positive by the classifier, True Negative (TN): Actual negative instances which are classified as negative by the classifier, False Positive (FP): Actual negative instances which are classified as positive instances by the classifier and False Negative (FN): Actual positive instances which are classified as negative instances by the classifier. The accuracy can be given below as eq (i),

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \text{ --------- (i)}$$

AUC is the arithmetic mean of TP Rate and TN rate for only one run of the classifier. When there are multiple runs of the classifier, AUC is the captured area in the Receiver Operative Curve (ROC) of TP Rate and TN rate for multiple runs. Another important measure used in Root Mean Square (RMS) Error.

5. Results

In this section, we presented the completed set of experimental observations. The proposed approach (Under Sampled using Random Forest) USRF is predominant on all the evaluation metrics. The accuracy, AUC and RMS Error are generated using 10 fold cross validation method.

Table 2. Accuracy on All the Datasets with Summary of Tenfold Cross Validation Performance

Dataset	C4.5	REP Tree	CART	NB Tree	ID3	USRF
Breast-cancer	74.28±6.05○	69.35±5.34●	70.22±5.19○	70.99±7.94○	58.95±9.22●	69.86±7.96
Breast-cancer-w	95.01±2.73●	94.79±2.74●	94.74±2.60●	96.38±2.23●	90.62± 3.20●	98.95± 1.22
Horse-colic	85.16±5.91○	84.94±5.73○	85.37±5.41○	81.71±6.39●	52.58± 8.09●	82.00± 7.71
German_credit	71.25±3.17○	72.02±3.38○	73.43±4.00○	74.27±4.22○	8.94± 3.03●	67.29± 4.54
Pima_diabetes	74.49±5.27●	74.46±4.39●	74.56±5.01●	75.24±5.23●	26.15± 4.31●	81.24± 4.57
Hepatitis	79.22±9.57●	78.62±7.07●	77.10±7.12●	80.93±9.66●	27.75±10.18●	82.54± 9.45
Ionosphere	89.74±4.38●	89.46±4.56●	88.87±4.84●	89.15±5.00●	17.32± 4.79●	94.49± 4.23
Labor	78.6±16.58●	78.2±17.09●	80.03±16.67●	92.27±11.79○	59.33±20.60●	87.10±14.47
Sick	98.72±0.55○	98.68±0.57○	98.85±0.54○	97.82±0.76	80.78±1.88●	97.82± 0.90
Sonar	73.61±9.34●	72.69±10.19●	70.72±9.43●	77.07±9.65●	70.96±1.93●	79.29± 10.42
Vote	96.57±2.56	95.33±3.10●	95.81±2.64●	95.03±3.29●	93.15±3.32●	96.24±3.0

Empty dot indicates the loss of USRF. ● Bold dot indicates the win of USRF;

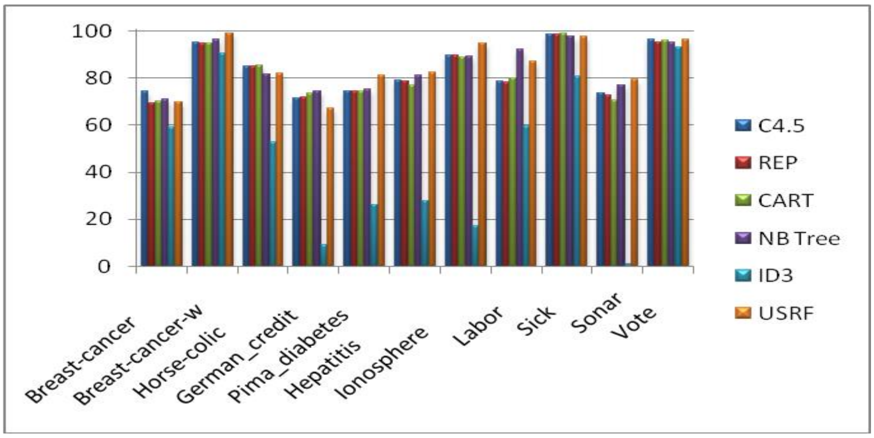
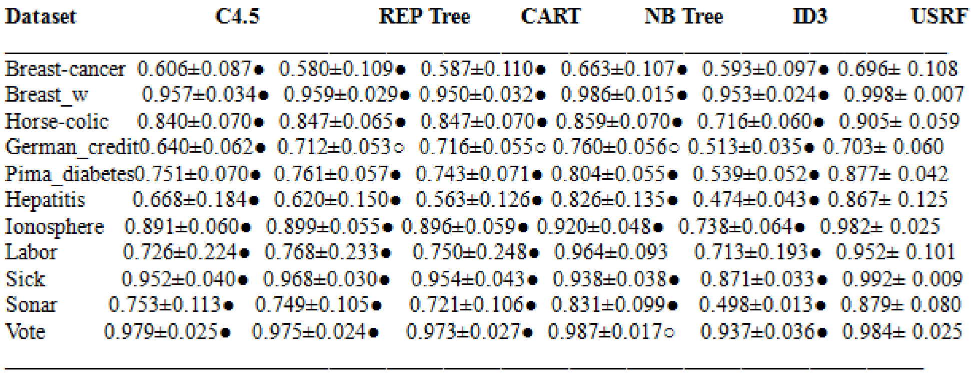


Figure 1. Trends in Accuracy for USRF versus Benchmark Algorithm on UCI Data Sets

Table 3. AUC on All the Datasets with Summary of Tenfold Cross Validation Performance



Dataset	C4.5	REP Tree	CART	NB Tree	ID3	USRF
Breast-cancer	0.606±0.087●	0.580±0.109●	0.587±0.110●	0.663±0.107●	0.593±0.097●	0.696± 0.108
Breast_w	0.957±0.034●	0.959±0.029●	0.950±0.032●	0.986±0.015●	0.953±0.024●	0.998± 0.007
Horse-colic	0.840±0.070●	0.847±0.065●	0.847±0.070●	0.859±0.070●	0.716±0.060●	0.905± 0.059
German_credit	0.640±0.062●	0.712±0.053○	0.716±0.055○	0.760±0.056○	0.513±0.035●	0.703± 0.060
Pima_diabetes	0.751±0.070●	0.761±0.057●	0.743±0.071●	0.804±0.055●	0.539±0.052●	0.877± 0.042
Hepatitis	0.668±0.184●	0.620±0.150●	0.563±0.126●	0.826±0.135●	0.474±0.043●	0.867± 0.125
Ionosphere	0.891±0.060●	0.899±0.055●	0.896±0.059●	0.920±0.048●	0.738±0.064●	0.982± 0.025
Labor	0.726±0.224●	0.768±0.233●	0.750±0.248●	0.964±0.093	0.713±0.193●	0.952± 0.101
Sick	0.952±0.040●	0.968±0.030●	0.954±0.043●	0.938±0.038●	0.871±0.033●	0.992± 0.009
Sonar	0.753±0.113●	0.749±0.105●	0.721±0.106●	0.831±0.099●	0.498±0.013●	0.879± 0.080
Vote	0.979±0.025●	0.975±0.024●	0.973±0.027●	0.987±0.017○	0.937±0.036●	0.984± 0.025

○ Empty dot indicates the loss of USRF. ● Bold dot indicates the win of USRF;

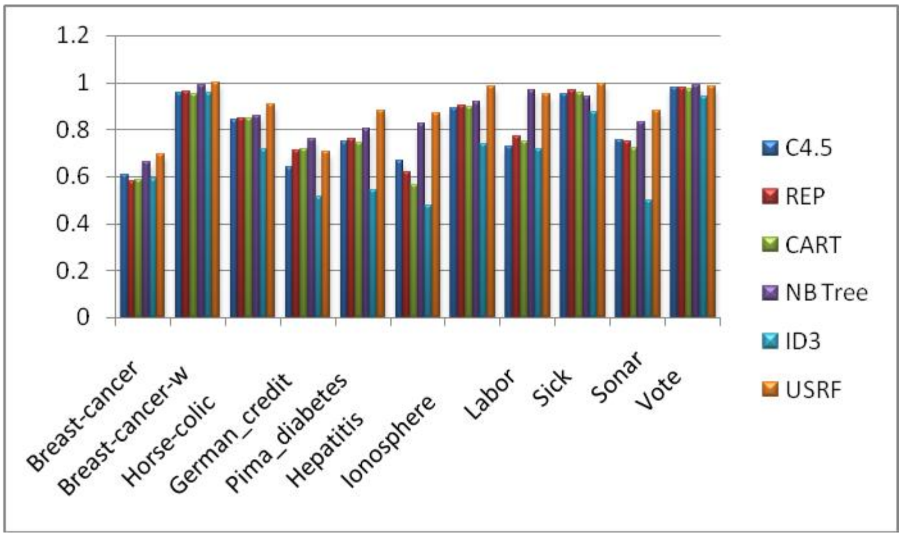


Figure 2. Trends in AUC for USRF versus Benchmark Algorithm on UCI Data Sets

Table 4. RMS Error on All the Datasets with Summary of Tenfold Cross Validation Performance

Dataset	C4.5	REP Tree	CART	NB Tree	ID3	USRF
Breast-cancer	0.444±0.037○	0.466±0.032●	0.458±0.039	0.473±0.057●	0.567±0.072●	0.458± 0.051
Breast_w	0.205±0.060●	0.209±0.056●	0.213±0.058●	0.169±0.062●	0.185±0.070●	0.087± 0.044
Horse-colic	0.352±0.060○	0.353±0.058○	0.346±0.059○	0.379±0.072●	0.391±0.105●	0.358± 0.048
German_credit	0.476±0.028●	0.441±0.025○	0.435±0.026○	0.428±0.034○	0.595±0.114●	0.458± 0.024
Pima_diabetes	0.439±0.042●	0.430±0.032●	0.432±0.036●	0.417±0.037●	0.624±0.059●	0.366± 0.037
Hepatitis	0.404±0.096●	0.402±0.057●	0.419±0.052●	0.371±0.099●	0.510±0.221●	0.344± 0.076
Ionosphere	0.299±0.081●	0.293±0.065●	0.302±0.068●	0.299±0.078●	0.050±0.131○	0.205± 0.053
Labor	0.401±0.170●	0.387±0.166●	0.380±0.183●	0.200±0.163○	0.425±0.274●	0.285±0.117
Sick	0.105±0.024○	0.106±0.023○	0.099±0.027○	0.136±0.024●	0.118±0.025○	0.127±0.018
Sonar	0.491±0.093●	0.452±0.071●	0.474±0.078●	0.434±0.098●	0.130±0.344○	0.374± 0.054
Vote	0.157±0.065○	0.186±0.061●	0.180±0.060●	0.185±0.068●	0.239±0.076 ●	0.172±0.063

○ Empty dot indicates the loss of USRF. ● Bold dot indicates the win of USRF;

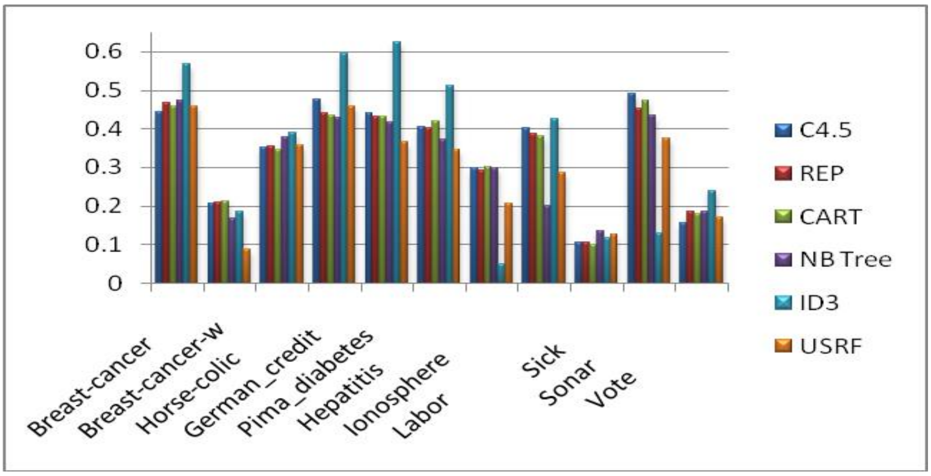
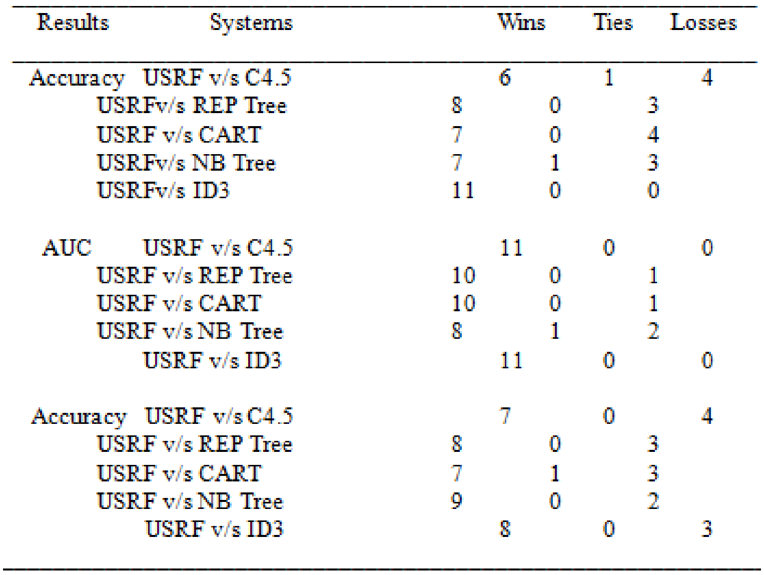


Figure 3. Trends in RMS Error for USRF versus Benchmark Algorithm on UCI Data Sets

From Table 2, we can observe that the proposed algorithms USRF accuracy value is improved on almost all the datasets. Table 3 compares the AUC value of the ID3 algorithm with the proposed USRF algorithm. The AUC value of the USRF algorithm is improved on all the datasets show that the USRF can handle the imbalance data efficiently.

Table 5. Summary of Experimental Results for USRF



Results	Systems	Wins	Ties	Losses
Accuracy	USRF v/s C4.5	6	1	4
	USRFv/s REP Tree	8	0	3
	USRF v/s CART	7	0	4
	USRFv/s NB Tree	7	1	3
	USRFv/s ID3	11	0	0
AUC	USRF v/s C4.5	11	0	0
	USRF v/s REP Tree	10	0	1
	USRF v/s CART	10	0	1
	USRF v/s NB Tree	8	1	2
	USRF v/s ID3	11	0	0
Accuracy	USRF v/s C4.5	7	0	4
	USRF v/s REP Tree	8	0	3
	USRF v/s CART	7	1	3
	USRF v/s NB Tree	9	0	2
	USRF v/s ID3	8	0	3

The results in Table 4 shows RMS error rate for the proposed USRF algorithm. The error rate of the USRF is decreased for all the UCI datasets. The fig 1, 2 and 3 represents the results of accuracy, AUC ad RMS error in the form of bar charts. The figures show that USRF has improved o all the metrics o almost all the UCI datasets against the compared C4.5, REP, CART, NB Tree and ID3 algorithms. The result in Table 5 presents the summary of the comparative study of USRF with the traditional algorithms.

6. Conclusion

We have proposed an effective and simple classification algorithm for handling class imbalance problem. This method uses the under sampling strategy which uses a unique way for identifying the surplus instances from the majority subset and balance the dataset to some extend. Empirical results have shown that USRF reasonable improved the results for reducing the imbalance effect compared with traditional methods. The proposed method can be extended to better visualize the unique properties of each and every datasets.

References

[1] 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.

[2] A. C. Lorena, L. F. O. Jacintho, M. F. Siqueira, R. D. Giovanni, L. G. Lohmann, A. C. P. L. F. de Carvalho and M. Yamamoto, "Comparing machine learning classifiers in potential distribution modeling", Expert Systems with Applications, vol. 38, **(2011)**, pp. 5268-5275.

[3] E. Molaei, H. Vadiatizadeh, Amirmahdimohammadighavam, N. Rajabpour and Fatemehziasistani "Distributed algorithm for privacy preserving data mining based on ID3 and improved secure sum", International Journal of Advanced studies in Computer Science and Engineering IJASCSE, vol. 3, no. 1, **(2014)**, pp. 28-34.

[4] Y. Hua, B. Feng, X. Z. Zhang, E. W. T. Ngai and M. Liu, "Stock trading rule discovery with an evolutionary trend following model", Expert Systems with Applications, vol. 42, **(2015)**, pp. 212-222.

[5] V. López, I. Triguero, C. J. Carmona, S. García and F. Herrera, "Addressing imbalanced classification with instance generation techniques: IPADE-ID", Neurocomputing, vol. 126, **(2014)**, pp. 15-28.

[6] S. Kumar and S. Jain, "Intrusion Detection and Classification Using Improved ID3 Algorithm of Data Mining", International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 5, **(2012)**, pp. 352-356.

[7] S. Manohar, A. Mittal, S. Naik and A. Ambre, "A Dynamic Classifier using Decision Tree Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, **(2015)**, pp. 628-631.

[8] N. Verbiest, E. Ramentol, C. Cornelisa and F. Herrerac, "Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection", Applied Soft Computing, vol. 22, **(2014)**, pp. 511-517.

[9] M. A. Hall, "Correlation-based feature subset selection for machine learning", PhD Thesis, **(1998)**.

[10] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, **(2001)**, pp. 5-32.

[11] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd edition Morgan Kaufmann, San Francisco, **(2005)**.

[12] J. Quinlan, "Induction of decision trees", Machine Learning, vol. 1, **(1986)**, pp. 81C106.

[13] L. Breiman, J. Friedman, R. Olshen and C. Stone, "Classification and Regression Trees", Belmont, CA: Wadsworth, **(1984)**.

[14] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid", In: Second International Conference on Knowledge Discovery and Data Mining, **(1996)**, pp. 202-207.

[15] A. Hamilton and A. 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)**.