International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(4), pp.156-164
DOI: http://dx.doi.org/10.21172/1.74.020
e-ISSN:2278-621X

AN IMPROVED APPROACH ON CLASS IMBALANCE DATA USING WITHIN-CLASS MINORITY OVERSAMPLING TECHNIQUE

D.Durga Prasad[1], Dr K. Nageswara Rao [2]

Abstract: Knowledge discovery from traditional or balance datasets can be done in an efficient way using the existing classification algorithms. The benchmark classification algorithms performance degrades when they are applied to the imbalance datasets. The reason is due to improper building of the predictive model using the imbalance datasets. In this paper, we propose a novel decision tree algorithm WithIn class Minority Oversampling TEchnique (WIMOTE) for efficient handling of imbalance data. The proposed WIMOTE approach uses the oversampling technique with unique statistical oversample strategy for removing misclassified and noisy instances in both majority and minority subset and oversamples the minority subset instances for data improvement. The experimental observation suggests that the proposed approach improves in terms of accuracy, AUC, Precision, Recall and F-measure with the benchmark SMOTE on 15 imbalance datasets from UCI repository.

Keywords: Data Mining, Knowledge Discovery, Classification, Decision Tree, imbalance data, WIMOTE

I. INTRODUCTION

Data mining is the process of discovering hidden or unknown knowledge from the existing datasets. The main approaches in data mining are supervised learning and unsupervised learning. In supervised learning, the data analyzed have the class labels and the classification is done I the predefined classes using the build model from the training data. In unsupervised learning, the analyzed data doest have any class labels. 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. It is noteworthy that class imbalance is emerging as an important issue in designing classifiers. Furthermore, the class with the lowest number of instances is usually the class of interest from the point of view of the learning task.

Whenever a class in a classification task is underrepresented (i.e., has a lower prior probability) compared to other classes, we consider the data as imbalanced. 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. In this

[1] *PP COMP SCI ENGG 431,Dept of CSE, Rayalaseema University, Kurnool(Dt), India.*
[2] *PSR & CMR College of Engineering, Vijayawada, India.*

paper, we propose a solution for the problem of class imbalance datasets using a novel oversampling strategy which uses the concept of with in class imbalance for efficient oversampling.

The arrangement of paper is follows as. We exhibit in Sec. 2 the recent approaches in decision tree learning. It will straight forwardly persuade the principle commitment of this work introduced in Sect. 3, somewhere we propose another structure for WIMOTE. Assessment criteria's designed for decision tree learning is exhibited in area 4. Test results are accounted for in Sect. 5. In conclusion, we finish up with Sect. 6 where we talk about real open issues and upcoming work.

II. CURRENT APPROACHES IN DECISION TREES

The decision tree approaches with imbalance data is presented by many of the researchers, one of the contribution is done by

Wacharasak Siriseriwan et al. [1] have proposed a Safe-Level SMOTE by generating synthetic instances away from possibly surrounding majority instances ad handling minority outcast with 1-nearest neighbour model. Han C et al. [2] have proposed novel online learning algorithms using passive-aggressive (PA) technique as well as a truncated gradient (TG) technique to solve high-dimensional imbalanced classification problem. L. Surya Prasanthi et al. [3] have proposed OSID3 approach using the oversampling technique with unique statistical oversample strategy for removing less privileged instances in the early stage and later on oversampling the high privileged instances for approximate data balance.

M Satya Srinivas et al. [4] have proposed an improved approach cost sensitive approach using Advanced Neuro Fuzzy Inference system (ANFIS). Win-Tsung Lo [5] et al., have designed and implement a new parallelized decision tree algorithm on a CUDA (compute unified device architecture), which is a GPGPU solution provided by NVIDIA where CPU is responsible for flow control while the GPU is responsible for computation. Andrea Dal Pozzolo [6] et al., have shown how Hellinger Distance Decision Trees can be successfully applied in unbalanced and evolving stream data by removing instance propagations between batches. Deepika Tiwari [7] has proposes a feature selection algorithm which is modified version of feature selection algorithm RELIEFF, this modifies the Original RELIEFF algorithm to handle the class imbalance problem by assigns higher weight to attributes while dealing with minority classes which results in higher weight of attributes which cater to minority samples.

Nicola Lunardon [8] et al., have proposed a ROSE package to deal with binary classification problems in the presence of imbalanced classes. Artificial balanced samples are generated according to a smoothed bootstrap approach and allow for aiding both the phases of estimation and accuracy evaluation of a binary classifier in the presence of a rare class. Bao-Gang Hu [9] et al., have investigates into cost behaviors of binary classification measures in a background of class-imbalanced problems using a new perspective for validation measures by revealing their cost functions with respect to the class imbalance ratio. Peng Cao [10] et al., have presents an effective wrapper approach incorporating the evaluation measure directly into the objective function of cost-sensitive neural network to improve the performance of classification, by

simultaneously optimizing the best pair of feature subset, intrinsic structure parameters and misclassification costs using Particle Swarm Optimization technique.

Hyoung-joo Lee[11] et al., have shown that the novelty detection approach is a viable solution to the class imbalance and examine which approach is suitable for different degrees of imbalance. They also applied SVM-based classifiers, when the imbalance is extreme, novelty detectors are more accurate than balanced and unbalanced binary classifiers. Giovanna Menardi [12] et al., have discussed the effects of class imbalance on model training and model assessing. A unified and systematic framework for dealing with both the problems is proposed, based on a smoothed bootstrap re-sampling technique. D.Ramyachitra [13] et al., have review differ imbalance approaches for class imbalance learning which are applicable in detection of fraudulent calls, bio-medical, engineering, remote-sensing, computer society and manufacturing industries.

III. THE PROPOSED APPROACH - WITHIN-CLASS IMBALANCE MINORITY OVERSAMPLING TECHNIQUE (WIMOTE)

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

Phase I: Preparation of the Majority and Minority subsets
The datasets is partitioned into majority and minority subsets. As we are concentrating on over sampling, we will take minority data subset for further visualization analysis to identify within-class imbalance assemblies.

Phase II: Initial phase of removing noisy and within-class assemblies' 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 those within-class assemblies 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 those particular within-class assemblies. This process can be applied on all the within-class assemblies identified for each dataset.

Phase III: Applying oversampling on within-class assemblies

The oversampling of the instances can be done on the improved within-class assemblies produced in the earlier phase. The oversampling can be done as follows:
Apply resampling supervised filter on the within-class assemblies 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 subset. Perform oversampling on within-class assemblies can help so as to form strong, efficient and more valuable rules for proper knowledge discovery.

Phase IV: Forming the strong dataset

The minority subset and 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 random forest [14] as the base algorithm.

The proposed WIMOTE algorithm is summarized as below.

WIMOTE algorithm

Input: A set of major subclass examples *P*, a set of minor subclass examples *N*, *jPj* < *jNj*, and *Fj*, the feature set, j > 0.

Output: Average Measure { accuracy, AUC, Precision }

Phase I: Initial Phase:
1: begin
2: $k \leftarrow 1$,j$\leftarrow$1.
3: Apply Visualization Technique on subset *N*,
4: Identify within-class assemblies *Cj from* N, j= number of within-class assemblies identified in visualization
5: repeat
6: k=k+1
7: Identify and remove the borderline and outlier instances for the within-class assemblies *Cj*.
8: Until k = j

Phase II: Over sampling Phase
9: Apply Oversampling on *Cj* within-class assemblies from N,
10: repeat
11: k=k+1
12: Generate '*Cj* × s' synthetic positive examples from the minority examples in each within-class assemblies *Cj*.
13: Until k = j

Phase III: Validating Phase
14: Train and Learn A Base Classifier (random forest) using *P* and *N*
15: end

IV. EXPERIMENTAL SETUP AND ASSESSMENT CRITERIA

Experiments are conducted using fifteen datasets from UCI [15] data repositories. Table 1 summarizes the benchmark datasets used in the anticipated study. For each data set, S.no., Dataset, name of the dataset, Instances, number of instances, Attributes, Number of Attributes, IR, Imbalance Ratio are described in the table for all the datasets.

Table 1 UCI datasets and their properties

S.no.	Dataset	Inst	Attributes	IR
1.	Breast	286	9	2.37
2.	Breast-cancer-w	699	9	1.90
3.	Horse-colic	368	22	1.71
4.	Credit-g	1000	20	2.33
5.	Pima diabetes	768	8	1.87
6.	Heart-c	303	13	1.19
7.	Heart-h	294	13	1.77
8.	Heart-statlog	270	14	1.25
9.	Hepatitis	155	20	3.85
10.	Ionosphere	351	35	1.79
11.	Kr-vs-kp	3196	37	1.09
12.	Labor	57	17	1.85
13.	Mushroom	8124	23	1.08
14.	Sick	3772	30	15.32
15.	Sonar	208	13	1.15

We performed the implementation of our new algorithms within the Weka [16] environment on windows 7 with i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM. The validation of the results is done using 10 fold cross validation, in which the dataset is split into 10 subsets and in each run nine subset are used for training and the remaining subset is used for testing. In 10 runs, the testing subset is altered and average measures for the 10 runs are generated. The evaluation metrics used in the paper are detailed below,

Accuracy is the percentage of correctly classified instances. AUC can be computed simple as the micro average of TP rate and TN rate when only single run is available from the clustering algorithm.
The Area under Curve (AUC) measure is computed by,

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

Or

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

The Precision measure is computed by,

$$\text{Pr}\,ecision = \frac{TP}{(TP) + (FP)} \quad \text{----------} \quad (3)$$

The Recall measure is computed by,

$$\text{Re}\,call = \frac{TP}{(TP) + (FN)} \quad \text{----------} \quad (4)$$

The F-measure Value is computed by,

$$F - measure = \frac{2 \times \text{Pr}\,ecision \times \text{Re}\,call}{\text{Pr}\,ecision + \text{Re}\,call} \quad \text{----------} \quad (5)$$

V. RESULTS

The experimental results of the proposed approach are presented in the below section. The proposed approach is compared with one of the popular synthetic minority oversampling algorithm, SMOTE [17]. Table 2 presents the results of AUC with the proposed approach WIMOTE. The proposed approach WIMOTE has win on 13 datasets and tie on 1 dataset and loss on 1 dataset. Table 3 presents the comparison of the precision results, in which the proposed approaches has win on 12 datasets and tie on 2 datasets and win on 1 dataset. Table 4 presents the comparison results of recall, in which the proposed approach win on 12 datasets, tie on 1 datasets and loss on 2 datasets. In terms of f-measure, the proposed approach has win on 13 datasets, tie on 1 dataset and loss on 1 dataset. Table 5 presents the comparison results of F-measure, in which the proposed approaches has win on 13 datasets and tie on 1 datasets and win on 1 dataset.

Table 2 Results of AUC on all the datasets with summary of tenfold cross validation performance

Datasets	SMOTE	WIMOTE
Breast	0.717±0.084●	0.900±0.062
Breast_w	0.967±0.025●	0.994±0.009
Colic	0.908±0.040●	0.976±0.020
Crerdit-g	0.778±0.041●	0.942±0.024
Diabetes	0.791±0.041●	0.939±0.021
Heart-c	0.830±0.077●	0.929±0.047
Heart-h	0.904±0.054●	0.961±0.033
Heart-s	0.832±0.062●	0.927±0.045
Hepatitis	0.798±0.112●	0.967±0.045
Ionosphere	0.904±0.053●	0.989±0.015
Kr-vs-kp	0.999±0.001○	0.998±0.002
Labor	0.833±0.127●	0.988±0.048
Mushroom	1.000±0.00	1.000±0.000
Sick	0.962±0.025●	0.998±0.004
Sonar	0.814±0.090●	0.955±0.051

● Bold dot indicates the win of Proposed approach;

Table 3 Results of precision on all the datasets with summary of tenfold cross validation performance

Datasets	SMOTE	WIMOTE
Breast	0.710±0.075●	0.895±0.063
Breast_w	0.974±0.025●	0.990±0.016
Colic	0.853±0.057●	0.955±0.041
Crerdit-g	0.768±0.034●	0.903±0.034
Diabetes	0.781±0.064●	0.895±0.041
Heart-c	0.779±0.082●	0.862±0.078
Heart-h	0.878±0.076●	0.945±0.054
Heart-s	0.791±0.081●	0.864±0.075
Hepatitis	0.709±0.165●	0.821±0.143
Ionosphere	0.934±0.049●	0.956±0.038
Kr-vs-kp	0.996±0.005○	0.984±0.009
Labor	0.871±0.151●	0.937±0.127

Mushroom	1.000±0.000	1.000±0.000
Sick	0.983±0.007●	0.995±0.004
Sonar	0.863±0.068	0.863±0.084

● Bold dot indicates the win of Proposed approach;

Table 4 Results of Recall on all the datasets with summary of tenfold cross validation performance

Datasets	SMOTE	WIMOTE
Breast	0.763±0.117●	0.825±0.084
Breast_w	0.947±0.035●	0.967±0.028
Colic	0.913±0.058○	0.904±0.067
Crerdit-g	0.810±0.058●	0.848±0.045
Diabetes	0.712±0.089●	0.832±0.054
Heart-c	0.777±0.110●	0.864±0.081
Heart-h	0.815±0.084●	0.861±0.080
Heart-s	0.803±0.110●	0.867±0.088
Hepatitis	0.681±0.188●	0.885±0.147
Ionosphere	0.881±0.071●	0.963±0.041
Kr-vs-kp	0.995±0.006○	0.994±0.006
Labor	0.765±0.194●	0.912±0.167
Mushroom	1.000±0.000	1.000±0.000
Sick	0.990±0.005●	0.996±0.003
Sonar	0.865±0.090●	0.911±0.091

● Bold dot indicates the win of Proposed approach;

Table 5 Results of F-measure on all the datasets with summary of tenfold cross validation performance

Datasets	SMOTE	WIMOTE
Breast	0.730±0.076●	0.856±0.060
Breast_w	0.960±0.022●	0.978±0.016
Colic	0.880±0.042●	0.927±0.042
Crerdit-g	0.787±0.034●	0.874±0.028
Diabetes	0.741±0.046●	0.861±0.037
Heart-c	0.772±0.070●	0.859±0.056
Heart-h	0.841±0.061●	0.898±0.053
Heart-s	0.791±0.072●	0.862±0.062
Hepatitis	0.677±0.138●	0.839±0.110
Ionosphere	0.905±0.048●	0.959±0.030
Kr-vs-kp	0.995±0.004○	0.989±0.005
Labor	0.793±0.132●	0.908±0.117
Mushroom	1.000±0.00	1.000±0.000
Sick	0.987±0.004●	0.996±0.003
Sonar	0.861±0.061●	0.883±0.070

● Bold dot indicates the win of Proposed approach;

The trends of accuracy and precision results are represented in the fig 1 and fig 2. The trends show that, the proposed approach has performed better than the compared SMOTE.

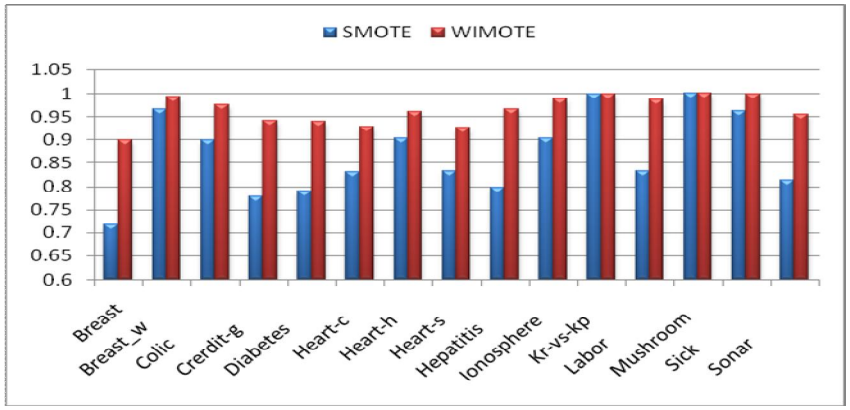

Fig. 1 Trends in AUC for WIMOTE versus SMOTE on imbalance data sets

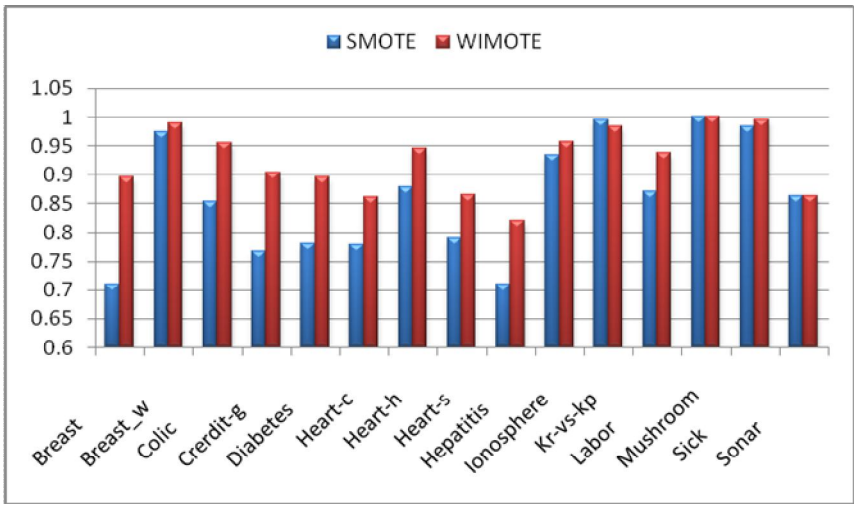

Fig. 2 Trends in Precision for WIMOTE versus SMOTE on imbalance data sets

VI. CONCLUSION

In this paper, we propose a novel decision tree algorithm WithIn class Minority Oversampling TEchnique (WIMOTE) for efficient handling of imbalance data. The proposed WIMOTE approach uses the oversampling technique with unique statistical oversample strategy for removing misclassified and noisy instances in both majority and minority subset and oversamples the minority subset instances for data improvement. The experimental observation suggests that the proposed approach improves in terms of accuracy, AUC, Precision, Recall and F-measure with the benchmark SMOTE on 15 imbalance datasets from UCI repository.

In future work, we will like to extend our system for high dimensional and complex datasets.

REFERENCES

[1] Wacharasak Siriseriwan and Krung Sinapiromsaran," The Effective Redistribution for Imbalance Dataset: Relocating Safe-eevel SMOTE with Minority Outcast Handling", Chiang Mai J. Sci. 2016; 43(1) : 234-246, http://epg.science.cmu.ac.th/ejournal/Contributed Paper.

[2] Han C, Tan YK, Zhu JH et al. Online feature selection of class imbalance via PA algorithm. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 31(4): 673–682 July 2016. DOI 10.1007/s11390-016-1656-0

[3] L. Surya Prasanthi, R. Kiran Kumar and Kudipudi Srinivas "An Improved ID3 Decision Tree Algorithm on Imbalance Datasets Using Strategic Oversampling", International Journal of Database Theory and Application Vol.9, No.5 (2016), pp.241-250 http://dx.doi.org/10.14257/ijdta.2016.9.5.25.

[4] M Satya Srinivas, A Yesubabu, G Pradeepini," Cost Sensitive Class Imbalance Learning using ANFIS", Australian Journal of Basic and Applied Sciences, 10(5) Special 2016, Pages: 144-149.

[5] Win-Tsung Lo, Yue-Shan Chang, Ruey-Kai Sheu, Chun-Chieh Chiu, Shyan-Ming Yuan,"CUDT: A CUDA Based Decision Tree Algorithm", Hindawi Publishing Corporation,The Scientific World Journal, Volume 2014, Article ID 745640, 12 pages, http://dx.doi.org/10.1155/2014/745640.

[6] Andrea Dal Pozzolo, Reid Johnson, Olivier Caelen, Serge Waterschoot, Nitesh V Chawla and Gianluca Bontempi," Using HDDT to avoid instances propagation in unbalanced and evolving data streams", In 2014 IEEE World Congress on Computational Intelligence (2014)

[7] DeepikaTiwari,"Handling Class Imbalance Problem Using Feature Selection", International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014) pp.no: 516-520, Vol. 2, Issue 2, Ver. 3 (April - June 2014).

[8] Nicola Lunardon, Giovanna Menardi, and Nicola Torelli," ROSE: A Package for Binary Imbalanced Learning", The R Journal Vol. 6/1, June, pp.no: 79-89.

[9] Bao-Gang Hu, Wei-Ming Dong,"A study on cost behaviors of binary classification measures in class-imbalanced problems", CoRR abs/1403.7100 (2014).

[10] Peng Cao , Dazhe Zhao and OsmarZaiane, "A PSO-based Cost-Sensitive Neural Network for Imbalanced Data Classification", Pacific-Asia Conference on Knowledge Discovery and Data Mining, 452-46.

[11] Hyoung-joo Lee and SungzoonCho,"The Novelty Detection Approach for Different Degrees of Class Imbalance", I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 21–30, 2006.Springer-Verlag Berlin Heidelberg 2006.

[12] GIOVANNA MENARDI, NICOLA TORELLI, "Training and assessing classification rules with unbalanced data", DEAMS working paper 2/2010.

[13] D. Ramyachitra, P. Manikandan, "IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW", International Journal of Computing and Business Research (IJCBR), ISSN (Online) : 2229-6166, Volume 5 Issue 4 July 2014

[14] Leo Breiman (2001). Random Forests. Machine Learning. 45(1):5-32.

[15] HamiltonA. 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

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

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