International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 10 (2017), pp. 2407-2417

http://www.ripublication.com

A Novel Technique on Class Imbalance Big Data using Analogous over Sampling Approach

Mohammad Imran
Regd No: PP.COMP.SCI&ENG.0308C, Research Scholar, Computer Science and Engineering, Rayalaseema University, Kurnool-518007, Andhra Pradesh, India ,

Dr.Vaddi Srinivasa Rao
Professor & Head Department of Computer Science and Engineering, Department of CSE, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada – 520007, Andhra Pradesh, India.

Amarasimha T
Research Scholar with Regd No: PP.COMP.SCI.0266 in the department of Computer Science,
Rayalaseema University, Kurnool-518007, Andhra Pradesh.

Syed Zeeshan Quadri
Pursuing M.S (CSE) in the Department of Computer Science and Engineering from university of Illinois, USA

Abstract

Big data consists of huge volumes of data which are used to analyses and discover the hidden knowledge. Class imbalance nature is a conventional issue which is present in all real world datasets. The class imbalance nature in the big data reduces the performance of the existing classification algorithms. In this paper, we propose a novel algorithm known as Over Sampling on Imbalance Big Data (OSIBD) which uses analogous oversampling strategy to improve the knowledge discovery from the class imbalance big datasets. The experimental simulations are conducted on eight moderately large class imbalance datasets which are obtained from UCI machine learning repository. The experimental results suggest that the proposed OSIBD algorithm had performed better than the existing C4.5 algorithm on class imbalance big datasets.

Index Terms — Classification, Imbalanced data, over sampling, OSIBD.

I. INTRODUCTION

Data mining is the process of discovering hidden knowledge from the existing databases. The main approaches are classification, clustering, association analysis and pattern mining etc.

Classification is the process of classifying the instances in the existing labeled classes by analyzing the features of the instances [1].The most popular classification techniques are decision trees, neural networks, support vector machines etc. In clustering, the instances are formed as clusters or groups depending upon the intrinsic properties of the instances. The popular clustering approaches are k-means, DB-scan, Hierarchical clustering etc.

In classification one of the effective and efficient approaches is decision trees. The decision trees are formed by the process of induction. The training instances are used to build the decision tree model and the testing subset is used to assess the performance of the build decision tree on the unseen instances. The class imbalance datasets are one of the new data source emerged recently. In binary class imbalance datasets, there exist two sub classes; majority and minority. The majority subclass is the one in which large percentage of instances from one class exists. In minority subset only less percentage of instances from other class exists. The performance of the existing classification drastically degrades when applied to class imbalance datasets. The reason for reduced performance is due to improper model built with the training instances. Since in the training subset, only few minority instances are available for model building. The model is very weak to predict the unseen minority instances.

The class imbalance problem also shows its presence in the case of big data sources in real time. In the context of big dataset the reduced performance is seen in the classification algorithms for class imbalance data. A new series of novel approaches are needed to address the problem of class imbalance on big data.

II. RELATED WORK

A. Existing problem and solution

Many algorithms and methods have been proposed to ameliorate the effect of class imbalance on the performance of learning algorithms. There are three main approaches to these methods.

• ***Internal approaches acting on the algorithm***. These approaches modify the learning algorithm to deal with the imbalance problem. They can adapt the decision threshold to create a bias toward the minority class or introduce costs in the learning

process to compensate the minority class.

• ***External approaches acting on the data*.** These algorithms act on the data instead of the learning method. They have the advantage of being independent from the classifier used. There are two basic approaches: oversampling the minority class and under-sampling the majority class.

Combined approaches that are based on boosting accounting for the imbalance in the training set. These methods modify the basic boosting method to account for minority class underrepresentation in the data set. There are two principal advantages of choosing sampling over cost-sensitive methods. First, sampling is more general as it does not depend on the possibility of adapting a certain algorithm to work with classification costs. Second, the learning algorithm is not modified, which can cause difficulties and add additional parameters to be tuned.

Peng Cao [2] 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 [3] 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 et al., [4] 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 et al., [5] 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. FRAMEWORK OF OSIBD ALGORITHM

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

Phase i: Preparation of the Majority and Minority subsets

The dataset is partitioned into majority and minority subsets. As we are concentrating over sampling, we will take minority subset for further visualization of class imbalance nature.

Phase ii: Improve with in class imbalances by removing noisy and 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 with in class 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 with in classes. This process can be applied in an analogous way to identified weak rages from the dataset.

Phase iii: Applying oversampling on the minority subset

The preprocessed minority subset is oversampled in an analogous approach by generating synthetic instances, replica instances and generating hybrid instances with the characteristics of existing and synthetic instances.

Phase iv: Forming the strong dataset

The improved minority subset and majority subset is combined to form a strong and reduced imbalance nature, which is used for learning with a base algorithm. In this case we have used C4.5 [6] as the base algorithm

.

IV. DATASETS

Experiments are conducted using eight datasets from UCI [7] 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 I. UCI DATASETS AND THEIR PROPERTIES

S.no.	Dataset	Inst	Attributes	IR
1.	Car	1728	7	18.61
2.	German_credit	1000	21	2.33
3.	Hypothyroid	3772	30	36.64
4.	Mfeat	2000	217	9.0
5.	Nursery	12960	9	13.17
6.	Page-blocks	5473	11	14.93
7.	Segment	2310	20	6.0
8.	Sick	3772	30	15.32

We performed the implementation of our new algorithms within the Weka [8] 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,

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

[Or]

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

The Precision measure is computed by,

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

The Recall measure is computed by,

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

V. EXPERIMETAL RESULTS

In the experimental setup, we have considered 8 datasets from UCI repository, which are in large size (few thousand instances). The validation is done by using the 10 fold cross validation for 10 runs. The mean of all the measures for 10 runs is used as experimental results. The reported experimental results suggest that our proposed OSIBD algorithm has performed well than the existing C4.5 algorithm. The C4.5 algorithm is one of the benchmark algorithms in decision trees.

The validation measures used in the experimental simulation are accuracy, AUC, Precision and Recall. In validation measures accuracy, AUC, Precision and Recall an increase in the values is to be reported for an improved performance.

If the proposed OSIBD algorithm is better than the compared technique then '●' symbol appears in the column. If the proposed OSIBD algorithm is not better than the compared technique then '○' symbol appears in the column. Table 2 reports the results of our proposed OSIBD algorithm verse C4.5 algorithm in terms of accuracy. The accuracy values generated by OSIBD algorithm are improved than C4.5 algorithm on 7 out of 8 datasets. Table 3 reports the results of our proposed OSIBD algorithm verse C4.5 algorithm in terms of AUC. The AUC values generated by OSIBD algorithm are improved than C4.5 algorithm on 5 out of 8 datasets. Table 4 reports the results of our proposed OSIBD algorithm verse C4.5 algorithm in terms of precision. The precision values generated by OSIBD algorithm are improved than C4.5 algorithm on 6 out of 8 datasets. Table 5 reports the results of our proposed OSIBD algorithm verse C4.5 algorithm in terms of recall. The recall values generated by OSIBD algorithm are improved than C4.5 algorithm on 4 out of 8 datasets. The mean performances were significantly different according to the T-test at the 95% confidence level

TABLE II. SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE FOR ACCURACY ON ALL THE DATASETS

Datasets	C4.5	OSIBD
Car	92.22±2.01●	92.50±1.86
German_credit	71.25±3.17●	77.02±3.60
Hypothyroid	99.54±0.36●	96.86±0.69
Mfeat	88.52±2.10●	89.94±1.95
Nursery	97.18±0.46○	70.96±1.10
Page-blocks	96.99±0.60●	97.10±0.60
Segment	96.79±1.29●	97.16±0.91
Sick	s98.72±0.55●	96.84±0.87

TABLE III. SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE FOR AUC ON ALL THE DATASETS

Datasets	C4.5	OSIBD
Car	0.981±0.011●	0.982±0.009
German_credit	0.647±0.062●	0.792±0.040
Hypothyroid	0.996±0.006○	0.984±0.011
Mfeat	0.979±0.024●	0.982±0.013
Nursery	1.000±0.000	1.000±0.000
Page-blocks	0.951±0.023●	0.958±0.025
Segment	0.994±0.009●	0.995±0.005
Sick	0.952±0.040○	0.943±0.036

TABLE IV. SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE FOR PRECISION ON ALL THE DATASETS

Datasets	C4.5	OSIBD
Car	0.972±0.016●	0.974±0.018
German_credit	0.767±0.025●	0.778±0.040
Hypothyroid	0.998±0.002●	0.988±0.008
Mfeat	0.934±0.055●	0.963±0.028
Nursery	1.000±0.000	1.000±0.000
Page-blocks	0.983±0.005●	0.984±0.005
Segment	0.984±0.025●	0.985±0.014
Sick	0.992±0.005○	0.977±0.008

TABLE V. SUMMARY OF TENFOLD CROSS VALIDATION PERFORMANCE FOR RECALL ON ALL THE DATASETS

Datasets	C4.5	OSIBD
Car	0.962±0.018●	0.963±0.016
German_credit	0.847±0.036○	0.806±0.054
Hypothyroid	0.998±0.002○	0.984±0.008
Mfeat	0.952±0.054●	0.961±0.027
Nursery	1.000±0.000	1.000±0.000
Page-blocks	0.986±0.005●	0.987±0.005
Segment	0.986±0.023●	0.993±0.013
Sick	0.995±0.004○	0.987±0.006

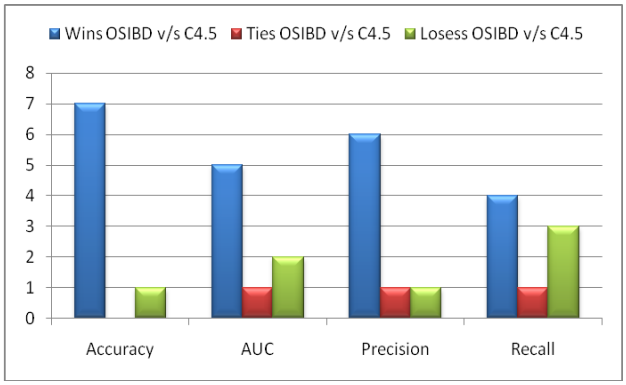

Figure 1. Trends of wins, ties and losses of OSIBD v/s C4.5 Unbalanced Big Datasets

TABLE VI. SUMMARY OF EXPERIMENTAL RESULTS FOR OSIBD

Results	**Systems**	**Wins**	**Ties**	**Losses**
Accuracy	OSIBD v/s C4.5	7	0	1
AUC	OSIBD v/s C4.5	5	1	2
Precision	OSIBD v/s C4.5	6	1	1
Recall	OSIBD v/s C4.5	4	1	3

Table 6 and figure 1 presents the summary of the experimental results of OSIBD algorithm verse C4.5 algorithm on different evaluation metrics. The registered more number of wins of OSIBD algorithm on C4.5 shows that our proposed algorithm is better than the existing algorithm on class imbalance datasets.

Finally, we can say that OSIBD is one of the best alternatives to handle class imbalance problems effectively. This experimental study supports the conclusion that the a prominent recursive over sampling approach can improve the CIL behavior when dealing with imbalanced datasets, as it has helped the OSIBD methods to be the best performing algorithms when compared with C4.5algorithm.

VI. CONCLUSION

Knowledge discovery from class imbalance big data in real world datasets is not efficiently performed using the existing classification algorithms. In this paper, we propose OSIBD algorithm using effective over sampling strategy for improved performance. The experimental results in terms of accuracy, AUC, precision and recall suggest that an improved performance can be achieved using the proposed algorithm.

REFERENCES

[1] O. Maimon, and L. Rokach, *Data mining and knowledge discovery handbook*, Berlin: Springer, 2010.

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

[3] Hyoung-joo Lee and Sungzoon Cho,”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.

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

[5] 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.

[6] J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann, 1993.

[7] Blake C, Merz CJ (2000) UCI repository of machine learning databases. Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html

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

AUTHORS BIOGRAPHY

Mohammad Imran received his B.Tech (CSE) and M.Tech (CSE) in 2008 from JNTU, Hyderabad, His Research interests include Big Data Analytics, Artificial Intelligence, Class Imbalance Learning, Ensemble learning, Machine Learning and Data mining. He is a Research Scholar with **Regd No: PP.COMP.SCI&ENG.0308C** in the department of Computer Science and Engineering, **Rayalaseema University**, **Kurnool-518007, Andhra Pradesh**. He is currently working as an Assistant Professor in Department of CSE, Muffakham Jah College of Engineering and Technology, Banjara Hills, Hyderabad-500034, India. You can reach him at imran.quba@gmail.com,

Dr.Vaddi Srinivasa Rao, Professor & Head Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada-520007,Andhra Pradesh, India, His Research Interest Includes Big Data Analytics, Computer Networks, Information Security, Artificial Intelligence, Class Imbalance Learning, Ensemble learning, Machine Learning and Data mining .

Amarasimha T received his M.C.A in 2004 from Visvesvaraya Technological University, Belagavi, Karnataka-590018. His Research interests include Big Data Analytics, Artificial Intelligence, Machine Learning and Data mining. He is a Research Scholar with **Regd No: PP.COMP.SCI.0266** in the department of Computer Science and Engineering, **Rayalaseema University**, **Kurnool-518007, Andhra Pradesh**.

Syed Zeeshan Quadri received his B.E (CSE) in 2016 from Osmania University,Hyderabad,India and is currently pursuing M.S (CSE) from university of Illinois, USA. His research interests include Machine Learning, Data mining, Big Data Analytics, Artificial intelligence, Cyber security.