A Novel Imbalance Learning with Fusion Sampling using Diversified Distribution

G. Shobana[1], Bhanu Prakash Battula[2]
[1]*Rajiv Gandhi University of Knowledge Technology, College in Nuzvid, India.*
[2]*Tirumala Engineering College , Narasaraopet, India.*

Abstract- In data mining, imbalance learning is a challenging task due to the intrinsic properties of the imbalance datasets. An imbalance data consists of unequal ratio instances in the classes. To address the limitations of imbalance data, we propose a novel algorithm dubbed as, Fusion Sampling using Diversified Distribution (FSDD) technique taking into account both under sampling and over sampling. In fact, our algorithm is capable of restructuring the original dataset at a very high conceptual level to alleviate the problems in the class imbalance. We conduct the empirical benchmark experimental setup using 12 datasets of varying class imbalance level. The experimental results suggest that the proposed approach performs effectively than the existing approaches.

Keywords- Data Mining, Knowledge Discovery, Classification, oversampling, under sampling, Fusion Sampling using Diversified Distribution (FSDD).

I. INTRODUCTION

Decision trees are the mathematical based algorithmic model which uses logic as the core unit for decision making. Decision tree consists of the branches and leaves. Each branch is a path of splitting the records into a narrow space and each leaf is the result of the classification of records in a specific class. There are numerous models of decision trees, which access the data and classify them in the predefined classes.

Rukshan Batuwita et al., [1] have reviewed to conclude that SVMs could produce suboptimal models which are biased towards the majority class and have low performance on the minority class. Rushi Longadge et al., [2] have gathered the evidence to show that the most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample when imbalance dataset are applied. Kun Jiang et al., [3] have propose a novel genetic algorithm-based SMOTE (GASMOTE) algorithm which uses different sampling rates for different minority class instances and finds the combination of optimal sampling rates.

Shaza M. Abd Elrahman et al., [4] have reviewed a general survey for class imbalance problem solutions and the most significant researcher‟s investigations. Bartosz Krawczyk [5] has provided a discussion and suggestions concerning lines of future research for classification, regression, clustering, data streams and big data analytics.

The review of the recent works suggests that the efficiency of the decision tree reduces drastically when applied for class imbalance data sources. The reason for the reduction in performance is due to the inefficient model built with the rare instances class.

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

II. RELATED WORK

Chongsheng Zhang et al [6] have reviewed a first empirical study on the performance of the two opposing pipelines for binary imbalance learning, i.e., first feature selection then resampling, or first resampling then feature selection. Shuo Wang et al [7] have performed a systematic study of handling concept drift in class-imbalanced data streams; including current research focuses and open challenges. Shuo Wang et al [8] have studied the issue of if and how class imbalance learning methods can benefit software defect prediction with the aim of finding better solutions. They investigated different types of class imbalance learning methods, including resampling techniques, threshold moving, and ensemble algorithms.

Lov Kumar et al [9] have conducted a study on the application of static source code metrics and machine learning techniques to predict aging related bugs in class imbalance software engineering datasets. Shuo Wang et al [10] have studied the combined challenges posed by multiclass imbalance and online learning, and aims at a more effective and adaptive solution. They introduced two resampling-based ensemble methods, called MOOB and MUOB, which can process multi-class data directly and strictly online with an adaptive sampling rate. M. Mostafizur Rahman et al [11] have examined the performance of over-sampling using SMOTE and an improved under-sampling technique to balance cardiovascular data.

Amritanshu Agrawal et al [12] have applied a multi-performance criteria‟s AUC and recall while fixing the weaker regions of the training data using SMOTUNED, which is an auto-tuning version of SMOTE. Jianhong Yan et al [13] have proposed a novel RE-sample and Cost-Sensitive Stacked Generalization (RECSG) method based on 2-layer learning models. The first step is Level 0 model

generalization including data pre-processing and base model training. The second step is Level 1 model generalization involving cost-sensitive classifier and logistic regression algorithm. Bo SUN et al [14] have introduce an under-sampling bagging framework and proposed an evolutionary under-sampling (EUS) based baggingensemble method EUS-Bag by designing a new fitness function considering three factors to make EUS better suited to the framework.

Sudarsun Santhiappan et al [15] have proposed a novel unsupervised topic modelling based weighting framework to estimate the latent data distribution using a topics oriented directed under-sampling algorithm that follows the estimated data distribution to draw samples from the dataset. Siqi Ren et al [16] have proposed an ensemble classifier called Gradual Resampling Ensemble (GRE), which handles data streams using a selectively re-sampling method, where drifting data can be avoidable, is applied to select a part of previous minority examples for amplifying the current minority set which exhibit concept drifts and class imbalance. Khaldy MA et al [17] have explored and analysed different feature selection methods to select a subset of the original data and then resample for a clinical dataset that suffersfrom high dimensional and imbalance data.

M. Muksitul Haque et al [18] have investigated a number of imbalanced class algorithms including the TAN+ AdaBoost algorithm for solving the imbalanced class distribution present in epigenetic datasets which inherently come with few differentially DNA methylated regions (DMR) and with a higher number of non-DMR sites. For this class imbalance problem, a number of algorithms are compared. Neelam Rout et al [19] have discussed the meaning of the imbalanced data, examples of the imbalanced data, different challenges of handling the imbalanced data, imbalance class problems and performance analysis metrics for the imbalanced data are elaborated in different scenario.

Georgios Douzas et al [20] have proposed a conditional version of Generative Adversarial Networks (cGAN) to approximate the true data distribution and generate data for the minority class of various imbalanced datasets to validate against multiple standard oversampling algorithms. Samir Al-Stouhi et al [21] have developed a method that is optimized to simultaneously augment the training data and induce balance into skewed datasets. Theyproposed a novel boosting-based instance transfer classifier with a label-dependent update mechanism that simultaneously compensates for class imbalance and incorporates samples from an auxiliary domain to improve classification. Brendan Juba et al [22] have consider themeasures of classifier performance in terms of precision and recall, a measure that is widely suggested as more appropriate to the classification of imbalanced data. They observed that whenever the precision is moderately large, the worse of the precision and recall is within a small constant factor of the accuracy weighted by the class imbalance ad the solution is that the only cure for class-imbalance is a larger number of examples.

III. THE METHOD ANTICIPATED

This section presents the detail architecture of the proposed Fusion Sampling using Diversified Distribution (FSDD) approach which consists of four major modules. The detailed working principles of the FSDD approach are explained below in the sub-sections.

In the initial stage of our frame work the dataset is divided into minority subset P € pi (i = 1,2,..., pnum) and majority subset N € ni (i = 1,2,..., nnum) respectively. The minority subset is the class of instances which are very less when compared to the other class in the dataset. The majority subset is the class of instances, which are more in percentage than the other class.

As the traditional algorithms efficiency drops down on imbalance data, to improve the efficiency, the dataset‟s majority subclass is to be under sampled or minority subclass is to be oversampled. In our proposed approach we initiated the both under sampling and oversampling strategy for the majority and minority sub classes respectively. One of the limitations of the existing oversampling algorithms is of not considering for removal of noisy and outlier instances before oversampling. Therefore, in the proposed approach before oversampling phase is started mostly misclassified instances are removed from the dataset in the form of under sampling. The technique proposed for identifying the mostly misclassified instances is by considering the nearest neighbor instances. If all the nearest neighbor instances of a particular instance are of opposite class then it implies that particular instance comes under the category of a noisy or outlier instance and can be eliminated.

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 [23] 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 eliminated instances can boost the performance of the proposed approach in two ways:

First it will reduce the noisy and outlier instances not only from majority but also minority subset and hence improves the quality of the dataset. Second it reduces some of the outlier and noisy instances from majority subset and so reduces the imbalance nature of the dataset.

In the next phase minority subset is oversampled. The some of the synthetic instances generated are the replica of the existing instances, hybrid instances and pure artificial instances.

In the final stage the fine tuned dataset is applied to base algorithm here random forest [24] is considered and evaluations metric are generated.

The algorithm for FSDD is elaborated below,

Fusion Sampling using Diversified Distribution (FSDD) 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, Recall, F-measure }

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

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

Phase III: Cluster reduction and Outlier removal
9: Apply Under sampling on*Cj*diversified distribution assemblies from P,
10: repeat
11: k=k+1
12: reduce sparse clusters identified using the nearest neighbor technique from the majority data space in each diversified distribution assemblies *Cj*.
13: Until k = j

Phase IV: Merging data spaces and Validation
14: Merge minority and improved majority data space
15: Train and Learn on a foundation Classifier (random forest) using *P* and *N*
16: end

IV. INVESTIGATIONAL DESIGN AND ASSESSMENT CRITERIA'S

The details of the datasets are given in table 1. 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. The most popular machine learning publicly available datasets are available at Irvine [25].

Table 1 The UCI datasets and their properties

S.no.	Dataset	Inst	Attributes	IR
1.	Breast	286	9	2.37
2.	Breast_w	699	9	1.90
3.	Colic	368	22	1.71
4.	Credit-g	1,000	20	2.33
5.	Diabetes	768	8	1.87
6.	Hepatitis	155	20	3.85
7.	Ionosphere	351	35	1.79
8.	Kr-vs-kp	3196	36	1.09
9.	Labor	57	17	1.85
10.	Mushroom	8124	22	1.07
11.	Sick	3772	30	15.32
12.	Sonar	208	13	1.15

The evaluation metrics used in the paper are detailed below, the percentage of instances correctly classified by a classified is known as Accuracy. 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 AUC is defined as the mean of true positive rate and true negative rate. The formula for AUC is given below,

$$AUC = \frac{1 + TPRATE - FPRATE}{2} \quad \ldots\ldots\ldots (1)$$

$$AUC = \frac{TPRATE}{2} \quad \ldots\ldots\ldots (2)$$

The Precision measure is computed by,

$$Precision = \frac{TP}{TP + FP} \quad \ldots\ldots\ldots\ldots (3)$$

The Recall measure is computed by,

$$Recall = \frac{TP}{TP + FN} \quad \ldots\ldots\ldots\ldots\ldots (4)$$

The F-measure value is computed by,

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \ldots (5)$$

V. RESULTS

In this section, the results of the FSDD approach is compared and discussed. In order to test the strength of our method compared to existing methods, we included Under Sampling and SMOTE in our experiments and the implementation of the proposed algorithm is done in the java programing language using the open source tool weka [26]. We evaluated each of the classifiers on the twelve datasets from a number of different resources of UCI data repositories (Table 1). The results are summarized as follows.

Table 2 Summary of tenfold cross validation performance for AUC on all the datasets

Datasets	RUS	SMOTE	FSDD
Breast	0.621±0.114●	0.717±0.084●	0.976±0.021
Breast_w	0.955±0.034●	0.967±0.025●	0.998±0.004
Colic	0.848±0.079●	0.908±0.040●	0.985±0.013
Credit-g	0.652±0.066●	0.778±0.041●	0.996±0.006
Diabetes	0.743±0.071●	0.791±0.041●	0.991±0.008
Hepatitis	0.725±0.214●	0.798±0.112●	0.986±0.029
Ionosphere	0.865±0.073●	0.904±0.053●	1.000±0.000
Kr-vs-kp	0.998±0.002	0.998±0.002	1.000±0.000
Labor	0.773±0.184●	0.833±0.127●	0.993±0.024
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.983±0.024●	0.962±0.025●	1.000±0.000
Sonar	0.731±0.118●	0.814±0.090●	0.995±0.010

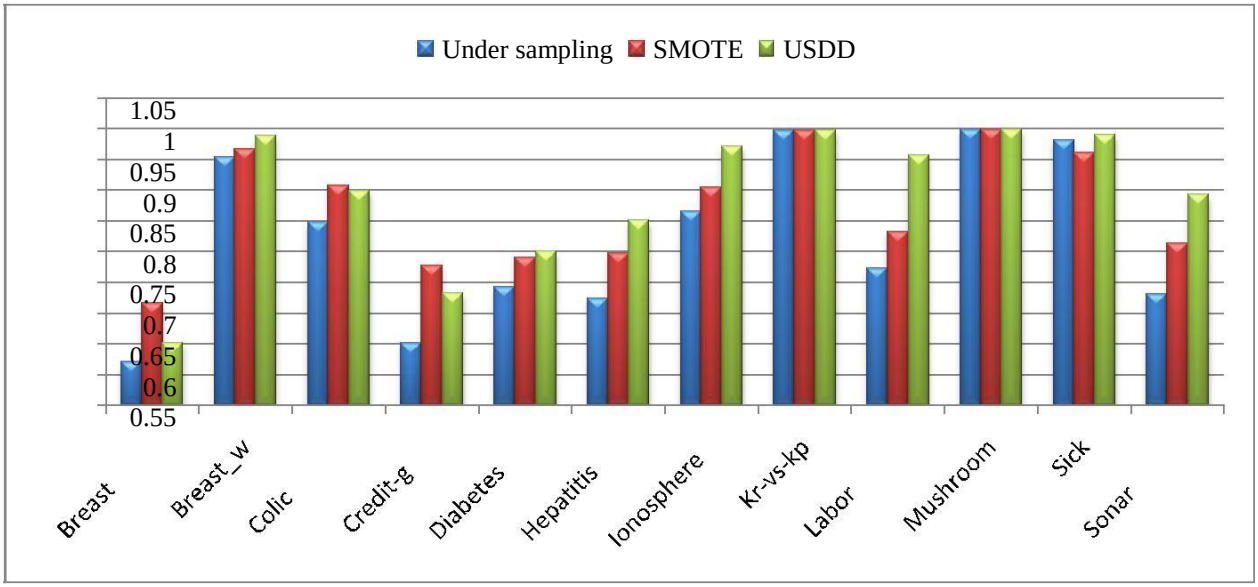

Fig.1: Trends in AUC for USDD versus RUS and SMOTE on UCI data sets

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

Datasets	RUS	SMOTE	FSDD
Breast	0.609±0.089●	0.710±0.075●	0.951±0.032
Breast_w	0.961±0.039●	0.974±0.025○	0.995±0.008
Colic	0.783±0.081●	0.853±0.057●	0.958±0.026
Credit-g	0.656±0.060●	0.768±0.034○	0.985±0.014
Diabetes	0.724±0.072●	0.781±0.064○	0.977±0.018
Hepatitis	0.692±0.268●	0.709±0.165○	0.960±0.081
Ionosphere	0.891±0.088●	0.934±0.049○	0.992±0.017
Kr-vs-kp	0.995±0.006○	0.975±0.006●	0.998±0.003
Labor	0.800±0.267●	0.871±0.151●	0.974±0.081
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.983±0.028●	0.983±0.007●	0.998±0.002

Sonar	0.711±0.110●	0.863±0.068○	0.982±0.036

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

Datasets	RUS	SMOTE	FSDD
Breast	0.730±0.146●	0.763±0.117●	0.956±0.040
Breast_w	0.940±0.040●	0.947±0.035●	0.995±0.009
Colic	0.917±0.082○	0.913±0.058○	0.970±0.028
Credit-g	0.661±0.094●	0.810±0.058●	0.982±0.014
Diabetes	0.720±0.102●	0.712±0.089●	0.979±0.019
Hepatitis	0.648±0.274○	0.681±0.188○	0.928±0.098
Ionosphere	0.844±0.102●	0.881±0.071●	0.991±0.020
Kr-vs-kp	0.994±0.007	0.994±0.007	1.000±0.002
Labor	0.765±0.297●	0.765±0.194●	0.948±0.109
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.968±0.042 ●	0.990±0.005●	1.000±0.000
Sonar	0.734±0.161●	0.865±0.090○	0.948±0.052

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

Datasets	RUS	SMOTE	FSDD
Breast	0.657±0.093●	0.730±0.076●	0.953±0.026
Breast_w	0.949±0.027●	0.960±0.022●	0.995±0.006
Colic	0.841±0.060●	0.880±0.042	0.964±0.019
Credit-g	0.656±0.065●	0.787±0.034●	0.983±0.010
Diabetes	0.718±0.071●	0.741±0.046●	0.978±0.013
Hepatitis	0.646±0.236○	0.677±0.138○	0.939±0.070
Ionosphere	0.861±0.064●	0.905±0.048○	0.991±0.013
Kr-vs-kp	0.994±0.004○	0.974±0.004●	0.999±0.002
Labor	0.750±0.237●	0.793±0.132●	0.955±0.078
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.975±0.028 ●	0.987±0.004●	0.999±0.001
Sonar	0.715±0.117●	0.861±0.061○	0.964±0.034

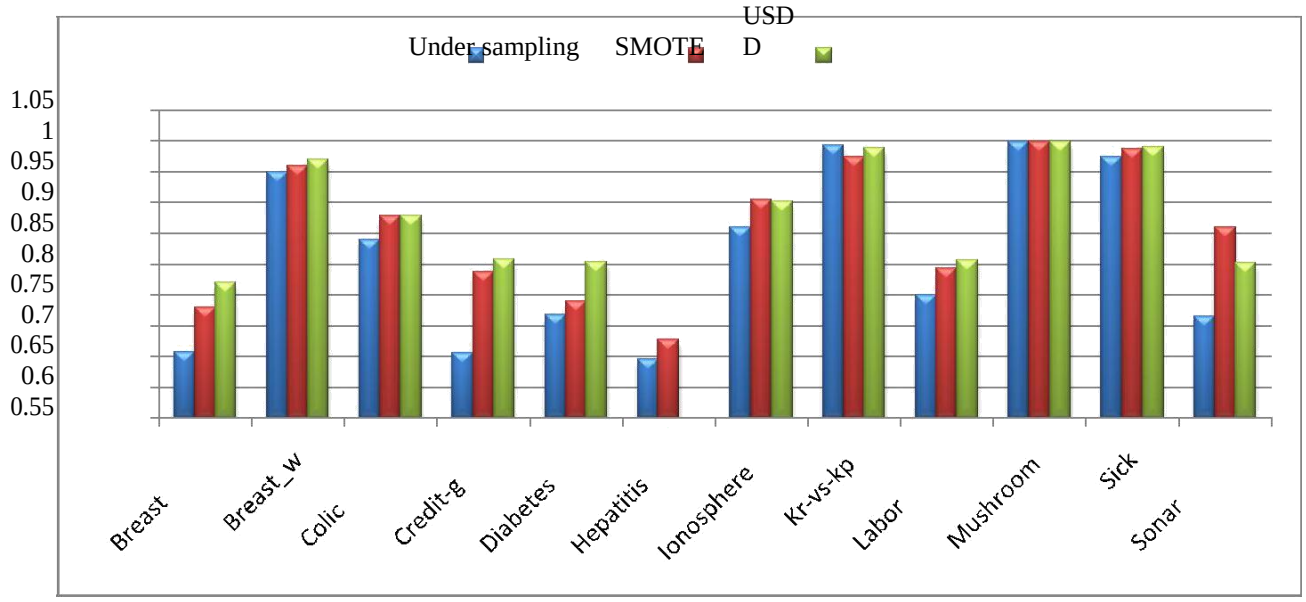

Fig.2: Trends in F-measure for USDD versus RUS and SMOTE on UCI data sets

Table 2 shows the detailed experimental results of the mean AUC of RUS, SMOTE on all the data sets verses proposed approach FSDD. From Table 2 we can see AUC performance of FSDD model with a substantial improvement over RUS on most data set (10 wins and 2 ties) which suggests that the FSDD model is potentially a good technique for decision trees. The FSDD method can also gain significantly improvement over SMOTE (10 wins and 2 ties) and is comparable to two state-of-the-art technique for decision trees.

Table 3 shows the detailed experimental results of the precision of RUS, SMOTE on all the data sets verses proposed approach FSDD. From Table 3 we can see FSDD model have performed well in terms of precision and have achieve substantial improvement over RUS, and SMOTE moderate improvement over FSDD.

Table 4 shows the detailed experimental results of the recall of RUS, SMOTE on all the data sets verses proposed approach FSDD. From Table 4 we can see error reduction of FSDD model with a substantial decrease over RUS on most data set (8 wins, 2 ties and 2 losses) which suggests that the FSDD model is potentially a good technique for decision trees. The FSDD method have reduced error over SMOTE (7 wins, 1 tie and 4 losses).

Table 5 shows the detailed experimental results of the mean F-measure of RUS, SMOTE on all the data sets verses proposed approach FSDD. From Table 5 we can see F-measure performance of FSDD model with a substantial improvement over RUS on most data set (9 wins, 1 ties and 2 losses) which suggests that the FSDD model is potentially a good technique for decision trees. The FSDD method can also gain significantly improvement over SMOTE (7 wins, 2 ties and 3 losses) and is comparable to two state-of-the-art technique for decision trees.

VI. CONCLUSION

In this paper, we propose a novel algorithm dubbed as, Fusion Sampling using Diversified Distribution (FSDD) technique taking into account both under sampling and over sampling.

In fact, our algorithm is capable of restructuring the original dataset at a very high conceptual level to alleviate the problems in the class imbalance. We conduct the empirical benchmark experimental setup using 12 datasets of varying class imbalance level. The experimental results suggest that the proposed approach performs effectively than the existing approaches.

In future work, we want to apply the proposed framework for multi class learning data sources.

VII. REFERENCE

[1]. Rukshan Batuwita and Vasile Palade,"CLASS IMBALANCE LEARNING METHODS FOR SUPPORT VECTOR MACHINES", Imbalanced Learning: Foundations, Algorithms, and
a. Applications,. By Haibo He and Yunqian Ma, Copyright c 2012 John Wiley & Sons, Inc.

[2]. Rushi Longadge, Snehlata S. Dongre, Latesh Malik," Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN) Volume 2, Issue 1, February 2013 www.ijcsn.org ISSN 2277-5420.

[3]. Kun Jiang, Jing Lu, Kuiliang Xia," A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE", Arab J Sci Eng, DOI 10.1007/s13369-016-2179-2.

[4]. Shaza M. Abd Elrahman and Ajith Abraham,"A Review of Class Imbalance Problem"Journal of
a. Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 332-340 © MIR Labs, www.mirlabs.net/jnic/index.html

[5]. Bartosz Krawczyk," Learning from imbalanced data: open challenges and future directions", Prog Artif Intell, DOI

10.1007/s13748-016-0094-0

[6]. Chongsheng Zhang, Jingjun Bi, Paolo Soda," Feature Selection and Resampling in Class Imbalance Learning: Which Comes First? An Empirical Study in the Biological Domain", 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp.o: 933-938.

[7]. Shuo Wang , Leandro L. Minku and Xin Yao," A Systematic Study of Online Class Imbalance Learning With Concept Drift", IEEE TRANSACTIONS ON NEURAL NETWORKS AND

a. LEARNING SYSTEMS.

[8]. Shuo Wang and Xin Yao," Using Class Imbalance Learning for Software Defect Prediction, IEEE

a. TRANSACTIONS ON RELIABILITY, VOL. 62, NO. 2, JUNE 2013.

[9]. Lov Kumar, Ashish Sureka," Feature Selection Techniques to Counter Class Imbalance Problem for Aging Related Bug Prediction", ISEC ‟18, February 9–11, 2018, Hyderabad, India.

[10]. Shuo Wang, Leandro L. Minku Xin Yao," Dealing with Multiple Classes in Online Class Imbalance

a. Learning", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16).

[11]. M. Mostafizur Rahman and D. N. Davis," Addressing the Class Imbalance Problem in Medical Datasets", *International Journal of Machine Learning and Computing, Vol. 3, No. 2, April 2013.*

[12]. Amritanshu Agrawal, Tim Menzies," Is "Better Data" Better Than "Better Data Miners"? On the Benefits of Tuning SMOTE for Defect Prediction", ICSE ‟18, May 27-June 3, 2018, Gothenburg, Sweden.

[13]. Jianhong Yan and Suqing Han," Classifying Imbalanced Data Sets by a Novel RE-Sample and Cost-

a. Sensitive Stacked Generalization Method", Mathematical Problems in Engineering Volume 2018, Article ID 5036710, 13 pages, https://doi.org/10.1155/2018/5036710.

[14]. Bo SUN, Haiyan CHEN, Jiandong WANG and Hua XIE," Evolutionary under-sampling based bagging ensemble method for imbalanced data classification", Front. Comput. Sci. DOI 10.1007/s11704-016-5306-z

[15]. Sudarsun Santhiappan, Jeshuren Chelladurai, and Balaraman Ravindran," A novel topic modeling based weighting framework for class imbalance learning", CoDS-COMAD ‟18, January 11–13, 2018, Goa, India

[16]. Siqi Ren, Bo Liao, Wen Zhu, Zeng Li, Wei Liu, Keqin Li," The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift", https://doi.org/10.1016/j.neucom.2018.01.063

[17]. Khaldy MA, Kambhampati C (2018) Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset. Int Rob Auto J 4(1): 00090. DOI: 10.15406/iratj.2018.04.00090