International Journal of Technology 10(4): 721-730
ISSN 2086-9614

DISTINCT MULTIPLE LEARNER-BASED ENSEMBLE SMOTEBAGGING (ML-ESB) METHOD FOR CLASSIFICATION OF BINARY CLASS IMBALANCE PROBLEMS

Dilip Singh Sisodia[1*], Upasna Verma[2]

[1]*Department of Computer Science and Engineering, National Institute of Technology Raipur, Raipur, Raipur, Chhattisgarh 492010, India*
[2]*Department of Information Technology, National Institute of Technology Raipur, Raipur, Chhattisgarh 492010, India*

(Received: March 2018 / Revised: February 2019 / Accepted: May 2019)

ABSTRACT

Traditional classification algorithms often fail in learning from highly imbalanced datasets because the training involves most of the samples from majority class compared to the other existing minority class. In this paper, a Multiple Learners-based Ensemble SMOTEBagging (ML-ESB) technique is proposed. The ML-ESB algorithm is a modified SMOTEBagging technique in which the ensemble of multiple instances of the single learner is replaced by multiple distinct classifiers. The proposed ML-ESB is designed for handling only the binary class imbalance problem. In ML-ESB the ensembles of multiple distinct classifiers include Naïve Bays, Support Vector Machine, Logistic Regression and Decision Tree (C4.5) is used. The performance of ML-ESB is evaluated based on six binary imbalanced benchmark datasets using evaluation measures such as specificity, sensitivity, and area under receiver operating curve. The obtained results are compared with those of SMOTEBagging, SMOTEBoost, and cost-sensitive MCS algorithms with different imbalance ratios (IR). The ML-ESB algorithm outperformed other existing methods on four datasets with high dimensions and class IR, whereas it showed moderate performance on the remaining two low dimensions and small IR value datasets.

Keywords: An ensemble of classifiers; Area under receiver operating curve; Classification; Class imbalance problem; Sensitivity; SMOTE; SMOTEBagging; SMOTEBoost; Specificity

1. INTRODUCTION

The advancement in data generation and acquisition tools and techniques has accelerated the growth and accessibility of raw data. This has further resulted in new avenues of learning from historical data (Sisodia et al., 2018b). The existing machine learning algorithms show good performance for many real-world applications with proportionate class instances. However, in the case of disproportionate instances (or imbalanced problems), the same learning algorithms face performance-related challenges. Therefore, in recent years, learning from imbalanced datasets has garnered significant attention of the machine learning community (Sun et al. 2018). In the datasets with binary classes, the majority class having more samples than the minority class overshadow the entire dataset (Collell et al., 2018). This class imbalance is further aggravated in critical real-world problems that have high misclassification cost of the minority

*Corresponding author's email: dssisodia.cs@nitrr.ac.in, Tel. +91-94-25371030, Fax. +91-771-225460
Permalink/DOI: https://dx.doi.org/10.14716/ijtech.v10i4.1743

class instances (Galar et al., 2012). Examples of such problems are determination of an uncommon disease, (Nusantara et al., 2016), fraud detection (Oentaryo et al., 2014; Moepya et al., 2014), bankruptcy prediction (Fedorova et al., 2013), intrusion identification in remote sensor (Rodda & Erothi 2016), oil spilling (Kubat et al., 1998), etc. Therefore, in datasets with imbalanced instances, learning algorithms are unable to appropriately represent the class distribution characteristics of the dataset. As a result, they produce objectionable credibility across the class of the dataset. The different techniques to deal with imbalanced data can be categorized into three approaches: data-level approach, algorithmic approach, and cost-sensitive learning approach. The data-level approach involves pre-processing of data before taking it into further consideration. Some of these approaches include random oversampling, random undersampling, Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), ADASYN (He et al., 2008), etc. The algorithmic approach, also known as an ensemble of classifiers, is used to improve the accuracy of the classifier (Chawla et al., 2003) by combining multiple classifiers. This approach shows significantly improved performance as compared to single classifiers. The cost-sensitive learning approach considers either the data-level or algorithmic approaches or both of them.

The remaining texts of this paper are organized under the following sections. In section two, research works related to class imbalance are discussed in brief. In section three, the working of the proposed approach is discussed in detail. In section four, evaluation parameters used for performance measure and comparison of the proposed approach are described. Section five describes the dataset used in this study with experimental results and discussions. Lastly, the conclusion and future work are summarized in section six.

2. RELATED WORK

Data sampling techniques are used to handle imbalance issues because they can improve the quality and robustness of the learning algorithm. In 2002, the SMOTE approach was proposed to balance class distribution by generating synthetic data for minority class. Its performance was evaluated based on nine benchmark datasets using C4.5, ripper and naïve Bayes classifiers with different degrees of imbalance (Chawla et al., 2002). In 2005, the Borderline-SMOTE was introduced to generate synthetic data for borderline minority class samples. Its performance was evaluated using four datasets and then compared with those of SMOTE, random oversampling and C4.5 (Han et al., 2005). The SMOTE approach for solving the classification performance of weak learners in multi-class datasets was later improved using a new ensemble model SMOTEBagging (Wang & Yao, 2009). In an ensemble, multiple learners effectively solve the class imbalance problem (Galar et al., 2012) and improve the performance of individual learners as the imbalance problem cannot be solved by learners individually. The EUSBoost (Galar et al., 2013) is another ensemble technique in which evolutionary techniques are combined with the undersampling method and tested over 30 datasets from KEEL repository. Its performance was evaluated by making comparisons with the performances of boosting, bagging and hybrid-based approaches. With the help of SDSMOTE (improved SMOTE method based on support degree) (Li et al., 2014), the minority class samples can be selected to generate new samples and recognize high positive class ratio on the whole dataset as compared to SMOTE.

In another previous study (Abolkarlou et al., 2014), the ensemble approach was proposed to solve class imbalance problem with SMOTE on ten datasets. The obtained results were compared with those of SMOTEBagging and SMOTEBoost. Another study proposed a novel ensemble-based technique (Zhang et al., 2014) in which the modified SMOTE was combined with Bagging, and the experiments were conducted on five datasets by comparing the results with those of Bagging and SMOTEBagging. In 2014, a method based on cost-sensitive decision tree with feature space partitioning was introduced (Krawczyk et al., 2014), and the results were computed on different benchmark datasets with varying imbalance ratios (IR). The analysis of SMOTEBagging with logistic regression using credit scoring data revealed its higher degree of accuracy compared to a simple logistic algorithm (Hanifah et al., 2015). A new ensemble classification method using random undersampling and ROSE sampling under a boosting scheme RHSBoost was proposed to address the imbalance classification problem (Gong & Kim, 2017). A study proposed the variants of SMOTEBoost for imbalanced regression task and evaluated its performance using 30 datasets (Moniz et al., 2018). Studies have also evaluated the effect of different data sampling methods on learning performance of individual and ensemble models using highly skewed bankruptcy and credit card fraud datasets (Sisodia & Verma, 2018; Sisodia et al., 2018a). The above-discussed ensemble methods perform better on some skewed dataset with small class IR but show poor performance on datasets with high IR. Therefore, this paper proposes a modified ensemble SMOTEBagging algorithm in which the bagging of multiple instances of the single base learner is replaced with distinct multiple base classifiers for improving the binary class prediction performance.

3. METHODOLOGY

The working of the proposed ensemble classification method for handling highly skewed binary datasets is described in this section. The main idea behind the proposed method is to combine multiple classifiers (such as Naïve Bays, Support Vector Machine, Logistic Regression, and Decision Tree (C4.5)) instead of multiple instances of single learner into the SMOTEBagging technique. The proposed Multiple Learners-based Ensemble SMOTEBagging (ML-ESB) technique is derived from the ensemble approach where the oversampling method called SMOTE integrates with an ensemble of classifiers, as shown in Figure 1.

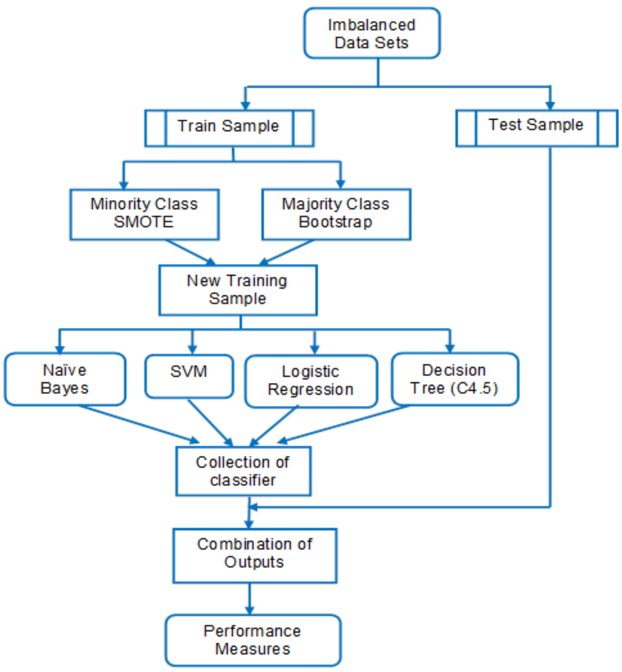

Figure 1 The visual illustration of working of the proposed methodology

SMOTE sampling technique is used to generate new synthetic data for minority class rather than sampling with replacement. This technique causes the classifier to build a large decision boundary for the nearby minority class. The working of ML-ESB is visually illustrated through a block diagram in Figure 1. The important steps used to explain the idea of ML-ESB algorithm are as follows. The pseudo-code used for implementation of ML-ESB is given as Algorithm1.

1. The dataset is divided into train and test samples using 10-fold cross-validation. From this training dataset, we derive the trained model. The training set consists of two classes: majority class having a large sample and minority class having less sample.
2. Now, this training set is bootstrapped with the majority class, and the minority class is oversampled using the sampling technique called SMOTE. By applying such an approach, we get a balanced training dataset with an equal number of minority and majority class samples.
3. The sampled training data is used to train the classifiers. The sampled training dataset is passed through the ensemble of classifiers such as Support Vector Machine, Naive Bayes, Logistic Regression and C4.5 to train the models.
4. The test samples are used to pass in the trained models, and the classification performance is ensemble using one of the popular majority voting techniques because it improves the performance of the classifier (Bauer & Kohavi, 1999).
5. The performance of ensemble of classifier is measured using G-mean, sensitivity, specificity, area under curve (AUC), and F-measure as these measures are preferable for imbalanced datasets.

Algorithm1 Pseudo code for ML-ESB algorithm

Input: Dataset having instances D = {1,2,3……N}

Output: Highest voted class label

Training the classifiers

Step1: Partition the dataset

a. For I = 1 to N
b. Divide each of the data into Target data as T and its class label as C separately.
c. Partition the data into K folds and find its equivalent $train_{data}$, $train_{class}$ and $test_{data}$, $test_{class}$

Step2: Construct subset D_k containing all the instances from both the classes with same number by executing the following steps:

a. Find the minority data O_d and class O_c =and majority data R_d =and its class R_c
b. **Resample the majority class**
   Resample all the training instances with replacement at 100% and generate a new majority class data R_d and its class label R_c
c. **Oversample, the minority class**
   Oversample the minority class instances using SMOTE (O_d, N, k)

Step3: Train the classifier from D_k

Test new instances

Step1: Pass new training instances to an ensemble of the classifier and generate output from each of the classifiers.

Step2: Return the class which gets the highest voting and then find all the measures.

The average computational complexity of the proposed ML-ESB is the same as SMOTEBagging for binary class imbalance problem because only ensemble of multiple instances of single classifier is replaced with distinct learners.

4. PERFORMANCE EVALUATION PARAMETERS

The classification accuracy is not a useful metric for evaluating the performance of learners because it gives the number of correct predictions from all predictions made. In the case of class imbalance problem, the classification accuracy completely ignores minority classes. Some other metrics including recall or sensitivity, specificity, and AUC are mostly used for evaluating the performance of learner for class imbalance problem. These metrics provide much greater insights into the performance characteristics of a classifier.

Specificity (Equation 1) is also called a true negative rate as it accurately measures the classified negative class.

$$\text{Specificity (SP)} = \frac{\text{TN}}{\text{FP+TN}} \tag{1}$$

Recall measures the completeness or sensitivity of a classifier. Higher recall means less false negatives, while lower recall means a false negative. The recall (Equation 2) is defined as the number of true positives over the number of true positives plus the number of false negatives.

$$\text{Sensitivity(SE)} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

where, TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, and TN is the number of true negatives.

The area under the receiver operating characteristic (AUC-ROC) (Huang & Ling, 2005) is used to evaluate the performance of the binary classifier. AUC is a two-dimensional curve plotted between sensitivity in Y-axis and 1-specificity in X-axis. The ROC curve is a relative trade-off between benefits and expenses that is useful for organizing the classifier and viewing the performance of the classifier specifically in the field of imbalanced distribution.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Dataset Description

To evaluate the performance of the proposed technique, we used six benchmark datasets from KEEL (Alcalá-Fdez et al. 2011) repository. All the datasets are highly imbalanced in nature with binary class labels. The brief descriptions of the datasets are given in Table 1.

Table 1 Descriptions of the datasets

Dataset	Brief Description	Number of Instances	Number of Attributes	Size (in KB)	IR
Pima	Pima is an Indian diabetes dataset; binary value in the diagnostics shows whether the patients are having diabetes according to the World Health Organization criteria.	768	8	23.6	1.87
Yeast	Yeast dataset is a real-world imbalanced dataset used to predict cellular localization sites of proteins.	1484	8	72.8	2.46
Vehicle	Vehicle dataset consists of vehicle attributes extracted from the silhouettes by the hierarchical image processing system for the classification of vehicles.	846	18	69.0	2.88
Segment	Segment data contains the information of segmented images drawn randomly from seven outdoor datasets of images.	2308	19	405	5.99
Page-blocks	The page-block dataset consists of block description of pages and is used for classifying the different blocks of pages such as text and graphic areas.	5472	10	286	8.79
Shuttle	Shuttle dataset is used for determining the type of control of the vessel that should be employed for the auto landing of a spacecraft.	1829	9	49.3	13.9

5.2. Results and Discussion

The extensive experiments were conducted using ML-ESB on six imbalanced binary datasets as described in Table 1. All the experiments were performed using a personal computer having 3.40GHz Core i7-4770 with 4.0 GB memory and running under the Microsoft Windows 8.1 Pro. The ML-ESB was implemented in MATLAB 2012 (MATLAB(2012a), n.d.). All experiments were run using 10-fold cross-validation. The results were evaluated using sensitivity, specificity and the AUC values. The respective ROC was also plotted for each dataset using ML-ESB. The performance of ML-ESB was compared with those of the other existing ensemble methods such as SMOTEBagging, SMOTEBoost and cost-sensitive MCS. The ML-ESB was evaluated with default IR as well as two different IR of 1:10 and 1:25.

5.2.1. Imbalance ratio 1:10

The first experiment was performed with ML-ESB on binary imbalanced datasets with IR 1:10. The same experiment was repeated with SMOTEBagging, SMOTEBoost and cost-sensitive MCS methods. The results were recorded based on two performance measures, i.e., sensitivity and specificity (Table 2). The results demonstrate the comparative performance of ML-ESB with the other existing methods. Table 2 shows that ML-ESB outperformed on four benchmark datasets, namely Vehicle, Segment, Page-blocks, and Shuttle. Acceptable results were obtained with two other datasets, i.e., Pima and Yeast.

Table 2 Results of the classifier with the imbalance ratio 1:10

Benchmark Datasets	SMOTEBagging		SMOTEBoost		Cost-Sensitive MCS		Proposed	
	SE	SP	SE	SP	SE	SP	SE	SP
Pima	84.01	96.12	84.01	96.12	85.23	97.10	91.35	66.60
Yeast	69.00	98.32	70.25	97.23	70.25	97.23	92.54	67.43
Vehicle	85.46	87.65	89.80	90.04	88.23	89.23	**98.18**	**94.91**
Segment	71.09	82.32	73.40	83.73	75.24	81.94	**99.69**	**99.39**
Page-blocks	77.43	77.89	77.98	79.34	82.95	80.23	**98.92**	**89.47**
Shuttle	87.56	90.23	89.54	90.23	92.31	89.23	**1.00**	**99.82**

The experimental results suggest that ML-ESB algorithm works better on four datasets with comparatively large numbers of features and high-class IR. It shows moderate results on the remaining two datasets with a small number of features and high IR.

Further, to confirm the demonstrated performance of ML-ESB, the AUC values were computed for all the six datasets with IR 1:10. The ML-ESB performance using AUC is reported in Table 3; it reflects the same trend as shown by sensitivity and specificity values. The AUC-ROC curve is also plotted for each dataset as shown in Figure 2.

Table 3 Result of the classifier with the imbalance ratio is 1:10

Dataset	AUC	Model Computation Time (Seconds)
Pima	68.97	6.60
Yeast	64.99	13.70
Vehicle	97.00	11.64
Segment	99.94	27.64
Page-blocks	94.20	77.60
Shuttle	99.99	11.69

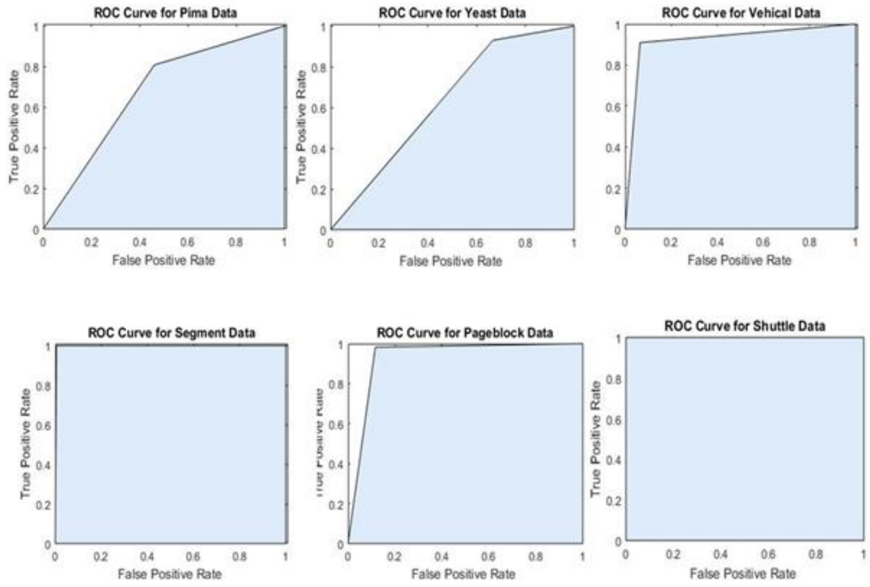

Figure 2 ROC plot for all the six datasets with IR 1:10

5.2.2. Imbalance ratio 1:25

The second experiment was performed with different values of IR of 1:25. The objective of this experiment is to ascertain the performance consistency of ML-ESB on different IRs. The same process of experimentation was adopted as done in the first experiment as discussed in previous subsection (5.2.1). The experimental results are recorded in Table 4, and it was observed that ML-ESB (for IR 1:25) outperformed on four benchmark datasets, namely Vehicle, Segment, Page-blocks, and Shuttle. However, it showed moderate performance on other two remaining datasets, i.e., Pima and yeast.

Table 4 Result of the classifier with the imbalance ratio of 1:25

Benchmark Datasets	SMOTEBagging		SMOTEBoost		Cost-Sensitive MCS		Proposed	
	SE	SP	SE	SP	SE	SP	SE	SP
Pima	75.21	92.32	75.43	94.35	85.23	97.10	95.14	65.66
Yeast	63.21	97.82	66.31	97.82	70.25	97.23	95.58	66.15
Vehicle	74.12	92.25	76.98	94.34	88.23	89.23	**97.2**	**94.89**
Segment	70.12	88.54	72.89	90.11	75.24	81.94	**99.69**	**99.39**
Page-blocks	71.25	83.65	73.89	82.97	82.95	80.23	**98.92**	**87.11**
Shuttle	83.10	90.58	85.02	91.89	92.31	89.23	**1.00**	**99.88**

Again, for IR 1:25, the performance of ML-ESB was evaluated using the AUC values and ROC plots for all benchmark datasets. Table 5 shows the AUC values, and the corresponding ROC plots are shown in Figure 3. The obtained results confirm the superior performance of ML-ESB, as shown by sensitivity and specificity values in Table 4.

Table 5 Result of the classifier with imbalance ratio of 1:25

Dataset	AUC	Model Computation Time (Seconds)
Pima	65.07	6.64
Yeast	60.87	13.65
Vehicle	96.05	11.78
Segment	99.54	27.58
Page-blocks	93.19	84.92
Shuttle	99.94	11.81

The experiment performed on benchmark datasets with IR 1:25 also demonstrated that ML-ESB algorithm works better on four datasets with comparatively large numbers of features and high-class IR. The proposed algorithm performed moderately on the remaining two datasets with a small number of features and high IR.

The model building time was also computed for using the proposed approach for both the experiments with IR 1:10 and 1:25 by using all datasets as shown in Tables 3 and 5. It is observed that the model construction time was proportional to the number of attributes in the datasets.

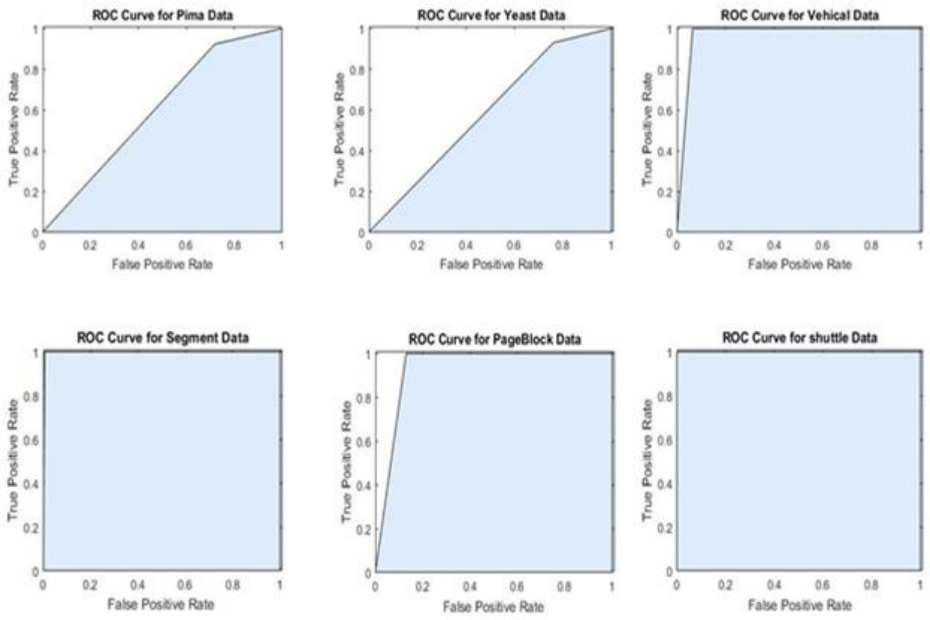

Figure 3 ROC plots for all the six datasets with IR 1:25

6. CONCLUSION

In this paper, a modified SMOTEBagging technique called ML-ESB is discussed to address the learning performance issues of imbalanced datasets. The performance of ML-ESB was evaluated on binary six benchmark imbalanced datasets using specificity and sensitivity with default class IR of datasets and two fixed IR values of 1:10 and 1:25 for all datasets. The obtained experimental results showed that the ML-ESB algorithm performed significantly better on four datasets with comparatively large numbers of features and high-class IR and performed moderately on the remaining two datasets with small number of attributes and low IR. In future, ML-ESB may be modified for handling multi-class imbalanced data classification.

7. REFERENCES

Abolkarlou, N.A., Niknafs, A.A., Ebrahimpour, M.K., 2014. Ensemble Imbalance Classification: Using Data Preprocessing, Clustering Algorithm and Genetic Algorithm. *In:* Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on. pp. 171–176

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., Garcia, S., Sánchez, L., Herrera, F., 2011. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing*, Volume 17(2–3), pp. 255–287

Bauer, E., Kohavi, R., 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, Volume 36(1-2), pp. 105–139

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, Volume 16, pp. 321–

357
Chawla, N.V., Lazarevic, A., Hall, L,O., Bowyer, K.W., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *In:* European Conference on Principles of Data Mining and Knowledge Discovery. pp. 107–119
Collell, G., Prelec, D., Patil, K.R., 2018. A Simple Plug-in Bagging Ensemble based on Threshold-moving for Classifying Binary and Multiclass Imbalanced Data. *Neurocomputing*, Volume 275, pp. 330–340
Fedorova, E., Gilenko, E., Dovzhenko, S., 2013. Bankruptcy Prediction for Russian Companies: Application of Combined Classifiers. *Expert Systems with Applications*, Volume 40(18), pp. 7285–7293
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-based Approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, Volume 42(4), pp. 463–484
Galar, M., Fernandez, A., Barrenechea, E., Herrera, F., 2013. EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-sets by Evolutionary Undersampling. *Pattern Recognition*, Volume 46(12), pp. 3460–3471
Gong, J., Kim, H., 2017. RHSBoost: Improving Classification Performance in Imbalance Data. *Computational Statistics and Data Analysis*, Volume 111, pp. 1–13
Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Advances in intelligent computing*, Volume 17(12), pp. 878–887
Hanifah, F.S., Wijayanto, H., Kurnia, A., 2015. SMOTE Bagging Algorithm for Imbalanced Dataset in Logistic Regression Analysis (Case : Credit of Bank X). *Applied Mathematical Sciences*, Volume 9(138), pp. 6857–6865
He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Proceedings of the International Joint Conference on Neural Networks*, (3), pp.1322–1328
Huang, J., Ling, C.X., 2005. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, Volume 17(3), pp. 299–310
Krawczyk, B., Woznaik, M., Schaefer, G., 2014. Cost-sensitive Decision Tree Ensembles for Effective Imbalanced Classification. *Applied Soft Computing Journal*, Volume 14(PART C), pp. 554–562
Kubat, M., Holte, R.C., Matwin, S., 1998. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, Volume 30(2–3), pp. 195–215
Li, K., Zhang, W., Lu, Q., Fang, X., 2014. An Improved SMOTE Imbalanced Data Classification Method based on Support Degree. *In:* International Conference on Identification, Information and Knowledge in the Internet of Things, pp. 34–38
MATLAB(2012a), Software package
Moepya, S.O., Akhoury, S.S., Nelwamondo, F.V., 2014. Applying Cost-sensitive Classification for Financial Fraud Detection under High Class-imbalance. *In*: 2014 IEEE International Conference on Data Mining Workshop, pp. 183–192
Moniz, N., Ribeiro, R.P., Cerqueira, V., Chawla, N.V., 2018. SMOTEBoost for Regression: Improving the Prediction of Extreme Values. In *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 150–159
Nusantara, A.C., Purwanti, E., Soelistiono, S., 2016. Classification of Digital Mammogram based on Nearest-Neighbor Method for Breast Cancer Detection. *International Journal of Technology*, Volume 7(1), pp. 71–77
Oentaryo, R., Lim, E-P., Finegold, M., Lo, D., Zhu, F., Phua, C., Cheu, E-Y., Yap, G-E., Sim, K., Nguyen, M.N., Perera, K., Neupane, B., Faisal, M., Aung, Z., Woon, W.L., Chen, W.,

Patel, D., Berrar, D., 2014. Detecting Click Fraud in Online Advertising : A Data Mining Approach. *Journal of Machine Learning Research*, Volume 15, pp. 99–140

Rodda, S., Erothi, U.S.R., 2016. Class Imbalance Problem in the Network Intrusion Detection Systems. *In:* 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 2685–2688

Sisodia, D.S., Reddy, N.K., Bhandari, S., 2018a. Performance Evaluation of Class Balancing Techniques for Credit Card Fraud Detection. *In:* IEEE International Conference on Power, Control, Signals and Instrumentation Engineering. IEEE, pp. 2747–2752

Sisodia, D.S., Singhal, R., Kandal, V., 2018b. Comparative Performance of Interestingness Measures to Identify Redundant and Non-informative Rules from Web Usage Data. *International Journal of Technology*, Volume 9(1), pp. 201–211

Sisodia, D.S., Verma, U., 2018. The Impact of Data Re-sampling on Learning Performance of Class Imbalanced Bankruptcy Prediction Models. *International Journal on Electrical Engineering and Informatics*, Volume 10(3), pp. 433–446

Sun, J., Lang, J., Fujita, H., Li, H., 2018. Imbalanced Enterprise Credit Evaluation with DTE-SBD: Decision Tree Ensemble based on SMOTE and Bagging with Differentiated Sampling Rates. *Information Sciences*, Volume 425, pp. 76–91

Wang, S., Yao, X., 2009. Diversity Analysis on Imbalanced Data Sets by using Ensemble Models. *In:* IEEE Symposium on Computational Intelligence and Data Mining. IEEE, pp. 324–331

Zhang, Y., Min, Z., Danling, Z., Gang, M., Daichuan, M., 2014. Improved SMOTEBagging and Its Application in Imbalanced Data Classification. *In:* 2013 IEEE Conference Anthology, ANTHOLOGY 2013