A Novel Approach to Solve Class Imbalance Problem Using Noise Filter Method

Gillala Rekha[1(✉)], Amit Kumar Tyagi[2], and V. Krishna Reddy[1]

[1] Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India
gillala.rekha@klh.edu.in
[2] Lingayas Vidyapeeth, Faridabad, Haryana, India

Abstract. Today's one of the popular pre-processing technique in handling class imbalance problems is over-sampling. It balances the datasets to achieve a high classification rate and also avoids the bias towards majority class samples. Over-sampling technique takes full minority samples in the training data into consideration while performing classification. But, the presence of some noise (in the minority samples and majority samples) may degrade the classification performance. Hence, this work introduces a noise filter over-sampling approach with Adaptive Boosting Algorithm (AdaBoost) for effective classification. This work evaluates the performance with the state of-the-art methods based on ensemble learning like AdaBoost, RUSBoost, SMOTEBoost on 14 imbalance binary class datasets with various Imbalance Ratios (IR). The experimental results show that our approach works as promising and effective for dealing with imbalanced datasets using metrics like F-Measure and AUC.

Keywords: Class imbalance · Ensemble learning method · Noise filter · Boosting · Over-sampling

1 Introduction

Most real-world classification problems often come across imbalanced datasets, i.e., fraud detection, identification, and diagnosis [4,11,19], event detection in water distribution system [15], diagnosis of disease [23], and face recognition [24], etc. This problem is known as class imbalance problems in classification, where the samples of some classes are extremely less than to other classes [12]. In particular, for a binary-class imbalance problem, the number of majority class samples proportion to the minority class samples, i.e., defined as 'an imbalance ratio'. From an application point of view, the minority class is always more interesting" [9] than majority class. In general most traditional classification algorithms,

Supported by KL University.

A. Abraham et al. (Eds.): ISDA 2018, AISC 940, pp. 486–496, 2020.
https://doi.org/10.1007/978-3-030-16657-1_45

for example, Decision Tree [2], Support Vector Machine [20], Bayesian Classification and Neural Network [10] face several difficulties in handling imbalanced datasets. These algorithms perform well on Balanced Class Distribution (BCD) but the learning algorithms show bias against an imbalanced dataset [13]. As a result, the traditional learning algorithms provide unfavourable accuracies across the classes of the data and tend to misclassify minority class samples. Such misclassification of minority class samples leads to several effects in real life problems. In general, nowadays sampling method is a widely used data pre-processing technique to deal with class imbalance problem. Sampling methods handle the imbalanced datasets by modifying the size of minority/majority class to provide a balanced distribution in a training dataset [9]. Recently, several researchers have proved that not all the samples are valued and donated to a classifier's learning [16,18,22]. Some samples may be redundant and tend to increase the computational cost. Some may even worsen a classifier's performance, which should be treated as noises and need to be removed/cleaned in both majority and minority classes. Thus, we intend to propose a framework to deal with the noisy examples in both minority and majority classes via a noise filter combined with over-sampling.

Hence, the organization of this work is followed as: Sect. 2 discusses the work related to class imbalance problem in brief. Our proposed method is discussed in Sect. 3 in detail. Then experiments or simulation results have been discussed in Sect. 4 with several parameters like AUC, F-Measure. Finally, Sect. 5 concludes this paper with some future enhancement in brief.

2 Related Work

According to the research which have done in [1,9,12], we find that under-sampling and oversampling methods are the popular and simple sampling methods for balancing the class distribution. In under-sampling, the majority class samples are discarded randomly to the size of minority samples and using over-sampling approach which generates synthetic data for minority samples to the size of majority samples. Both methods have been useful for handling class imbalance problems, but these approaches have several drawbacks. In under-sampling technique, the important majority samples may be ignored and which leads to a poor performance of the classifier. An oversampling technique tends to result in overfitting due to duplicate minority samples [6]. Apart under-sampling, the most popular oversampling technique, Synthetic Minority Over-sampling TEchnique (SMOTE) [3] has been proposed several authors [1,6]. SMOTE generates synthetic data based on the dimension space similarities among existing minority samples using nearest neighbour approach.

Apart that in the past decade, many SMOTE variants have been also proposed by several researchers, i.e., borderlineSMOTE [7], Safe-level-SMOTE [1], LN-SMOTE [14], ADASYN [8] in order to improve its performance. Using borderlineSMOTE the minority samples are over sampled only near the borderline instead of over-sampling all the minority samples. An improved

borderline-SMOTE is Safe-level-SMOTE algorithm, which uses a safe level to find the qualified points of new synthetic samples. To avoid the overgeneralization problem of SMOTE, LN-SMOTE defines a new safe level and rules to generate samples, and identify more exact information about the local neighborhood of its considered examples. ADASYN uses informative sampling approach instead generating samples of randomly. In the past decade, the implementation of class imbalance learning usually combines sampling with cost-sensitive methods, for example, AsymBoost [5] and AdaCost [6] and ensemble techniques, for example, AdaBoost [6] and Asymmetric bagging [21]. In summary, sampling techniques can play a great role in improving classifier accuracy and provide efficient results.

Recently, some researchers have evidenced that not all the samples are valuable and donated to a classifier's learning [22]. Some samples may be redundant and tend to increase the computational cost. Some may even worsen a classifier's performance, which should be treated as noises and need to be removed/cleaned in both majority and minority classes. However, to the best of our knowledge, there is not much work available to verify the influence of noise (available in both, i.e., in the minority/majority class). Thus, we intend to propose a framework to deal with the noisy examples in both minority and majority classes via a noise filter combined with over-sampling. Note that in this work, we focus only on binary classification problems. This represents the first attempt to combine the noise filter with re-sampling methods. In order to verify the efficiency, we choose three popular over-sampling methods, i.e., AdaBoost, RUSBoost, SMOTEBoost to implement the proposed framework with a K-Nearest Neighbor (KNN)-based noise filter. We design several experiments to test our proposed method with collected datasets (from KEEL Machine Learning Repository). In last, the propose framework is compared with the two basic methods, i.e., Area Under the Curve (AUC) and F-measure.

Hence, this section discusses about related work for handling class imbalance problem at data level. Now, next section deal with proposed data level framework for class imbalance problem.

3 Noise Filtering and Sampling Technique: A Proposed Framework

Noise filtering is a kind of pre-processing technique that are used to detect and remove noises in a dataset. It becomes basically needed since there are almost always noises presents in real-world datasets. It is also known that existence of such noises can weaken a classifier performance severely. Sáez et al. [17] proposed an over-sampling algorithm that uses an Iterative-Partitioning Filter (IPF) called, SMOTE-IPF to pre-process data. The training dataset is split into n subsets and are trained using a set of n base classifiers independently. A sample is identified as a noise one if it is misclassified by the base classifiers. Such samples are then removed from the training dataset before performing classification.

To the best of our knowledge, existing noise filters in the literature are always combined with under sampling, over-sampling techniques or only deal with the noisy examples in the majority class. No noise filtering attempt focuses on entire dataset using ensemble approach in the process of solving for imbalanced classification problem. Can such challenge boost a classifier's performance? Answering this question, this work proposes a Noise-filtered over-sampling technique with boosting (NF-OS with Boosting), as shown in Fig. 1.

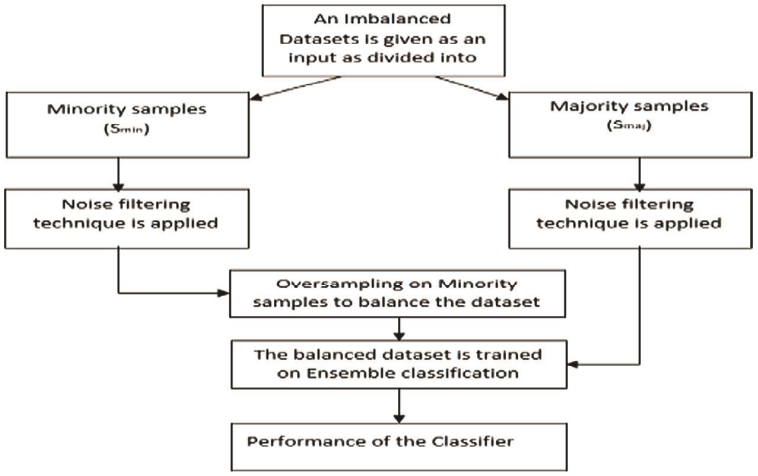

Fig. 1. A Noise-filtered Over-Sampling Technique with Boosting. (NF-OS with Boosting)

NF-OS with boosting is based on the combination of noise filtering with sampling and Adaboost algorithm. It is similar to RUSBoost and SMOTEBoost with the critical difference with removal of noise occurring in the datasets. SMOTEBoost uses SMOTE method to oversample the minority class examples, while RUSBoost uses random under-sampling on the majority class. In comparison, our proposed NF-OS uses noise filtering with sampling from the majority class. Considering a given dataset D,

I. We define subsets Smaj $\subset$ D and Smin $\subset$ D, where Smin is the set of minority samples in D, and Smaj is the majority class.

II. The noise in minority and majority samples are removed using K-Nearest Neighbours as follows (shown in the Fig. 2).

- [A] Each sample falls into any of the six categories based on nearest neighbours.
  - [a] Extremely useful majority sample: All the KNN are majority class label (labelled as 'A' in the Fig. 2).
  - [b] Extremely useful minority sample: All the KNN are minority class label (labelled as 'a' in the Fig. 2).
  - [c] Relatively useful majority sample: Most of the KNN samples belong to majority class label (labelled as 'B' in the Fig. 2).

−− [d] Relatively useful minority sample: Most of the KNN samples belong to minority class label (labelled as 'b' in the Fig. 2).
−− [e] Noisy sample: All the KNN belongs to different class label (both for majority and minority samples) (labelled as 'C' and 'c' in the Fig. 2).

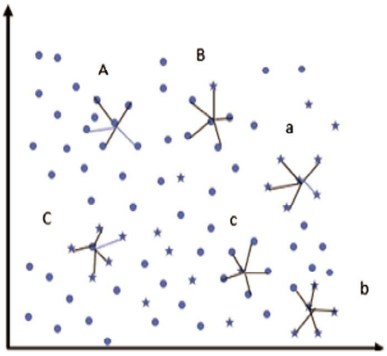

Fig. 2. Six categories of samples. '*' represents minority data. 'o' represents majority data

In this work, 'noisy samples' are identified using KNN algorithm. Note that the choice of K will be highly influenced to find whether a sample is a noise or not. If K is too small then a sample can be classified as a noise and if K is too large then it is considered as a useful one. The best value of K is this work is considered as 5. The strength of our approach lies in the fact that it considers examples after removal of noise in the entire dataset. After, NF-OS applies SMOTE technique to oversampling the minority samples in order to balance the imbalanced dataset. Once the dataset is balanced, classification is done using Boosting method. Note that, the Boosting algorithm considers a series of decision trees using C4.5 algorithm and combines the votes of each individual tree to classify new sample. Hence, this section discusses the proposed framework for handling class imbalance problem at data level.

Now, next section deal with the simulation and result for the proposed framework.

4 Experiments and Results

This section presents several experimental parameters used in this work and result with respect to proposed approach (noise filtering with Boosting). Here first, we present the evaluation metrics, benchmark datasets used and experimental settings for class imbalanced learning. Then, we show the results in form of two performance metrics for imbalanced learning, i.e., the AUC and F-measure.

4.1 Experimental Setting

In our experiment work, we tested the proposed scheme on 14 benchmark datasets shown in Table 2. For every dataset, we used C4.5 decision tree algorithms as a base learner in boosting. Here, each experiment is done with 20 independent runs with 10-fold cross validation and acquires the average results in terms of AUC and F-measure, respectively.

4.2 Evaluation Metrics

Traditionally, the commonly used metric for evaluating the performance of balanced classification algorithms is the error rate 'err' (refer Eq. 1). It is defined as 'percentage of number of misclassified samples by total number of samples'.

$$err = (number\,of\,misclassified\,samples) \,/\, (Total\,number\,of\,samples) \times 100 \quad (1)$$

However, it is not appropriate for classifying the imbalanced data sets. For example, there is a two-class imbalanced problem with an imbalanced rate of 99:1, with 99 majority samples and only 1 minority one. The goal of traditional learning algorithm is to minimize the error rate and for imbalanced data set with 99:1 rate, which may simply group all the samples into the majority class and thus attains 1% error rate. So, these learning algorithms are not a good approach to this problem. Since the only minority sample to which we should pay more attention is incorrectly classified. For this reason, different methods must be defined and used to validate the algorithms for handling class imbalance problems appropriately. In this paper, we study the two-class problems, in which the minority class is considered to be the positive class. Hence, the confusion matrix of a two-class problem shows the results of correctly and incorrectly classified samples of each class [5], as shown in Table 1.

Table 1. Confusion matrix

	Predicted positive class	Predicted negative class
Actual positive class	True Positive (TP)	False Negative (FN)
Actual negative class	False Positive (FP)	True Negative (TN)

In the literature, Receiver Operating Characteristics (ROC) curve evaluation metric has been proposed by many researchers for class imbalance problem. ROC makes use of the proportion of True Positive Rate (TPR) and False Positive Rate (FPR) (refer Eqs. 2 and 3).

$$TPR = TP \,/\, (TP + FN) \quad (2)$$

$$FPR = FP \,/\, (FP + TN) \quad (3)$$

The ROC graph is formed by plotting TPR over FPR, and any point in ROC space corresponds to the performance of a single classifier on a given distribution. The ROC curve is useful because it provides a visual representation of the relative trade-offs between the benefits (reflected by true positives) and costs (reflected by false positives) of classification with respect to data distributions [10]. Here, AUC is defined as the area under the ROC curve, which has been proved to be a reliable evaluation criterion and used as a metric to measure the efficiency against imbalanced classification problems. Two other important evaluation metrics [5] for imbalanced classification problems are defined as follows:

$$F - measure = (2 * precision * recall) / (precision + recall) \tag{4}$$

Where precision is defined as TP by TP and FP and recall is defined as TP by TP and FN.

4.3 Dataset Used

In this work, we test the proposed method on 14 benchmark datasets from KEEL-dataset repository with different imbalance ratio shown in Table 2.

Table 2. Data-set used

Datasets	Size	# attr	% IR
ecoli-0_vs_1	220	7	1.82
glass6	214	9	6.38
haberman	306	3	2.78
iris0	150	4	2
new-thyroid1	215	5	5.14
page-blocks0	5472	10	8.79
pima	768	8	1.87
segment0	2308	19	6.02
vehicle1	846	18	2.9
vehicle2	846	18	2.88
vehicle3	846	18	2.99
wisconsin	683	9	1.86
yeast1	1484	8	2.46
yeast3	1484	8	8.1

4.4 Results

Here, Table 3 shows the classification performance in terms of AUC obtained using different classification techniques. As indicated by the results, the proposed NF-OS with boosting demonstrated the best performance on 12 out of 14

Table 3. Classification performance using AUC metric

Datasets	AdaBoost	RUSBoost	SMOTEBoost	NF-OS Boost
ecoli-0_vs_1	0.6354	0.794	0.799	0.992
glass6	0.947	0.918	0.991	0.997
haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	1	0.99
new-thyroid1	0.947	0.975	0.947	0.986
page-blocks0	0.637	0.953	0.967	0.996
pima	0.6223	0.751	0.897	1
segment0	0.996	0.994	0.998	0.998
vehicle1	0.754	0.768	0.897	1
vehicle2	0.854	0.966	0.967	1
vehicle3	0.745	0.763	0.894	1
wisconsin	0.9	0.96	0.994	1
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994

Table 4. Classification performance using F-Measure metric

Datasets	AdaBoost	RUSBoost	SMOTEBoost	NF-OS Boost
ecoli-0_vs_1	0.6354	0.794	0.799	0.992
glass6	0.947	0.918	0.991	0.997
haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	1	0.99
new-thyroid1	0.947	0.975	0.947	0.986
page-blocks0	0.637	0.953	0.967	0.996
pima	0.6223	0.751	0.897	1
segment0	0.996	0.994	0.998	0.998
vehicle1	0.754	0.768	0.897	1
vehicle2	0.854	0.966	0.967	1
vehicle3	0.745	0.763	0.894	1
wisconsin	0.9	0.96	0.994	1
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994

datasets in terms of AUC for almost many datasets. Similarly, Table 4 shows the classification performance in terms of F-measure. The F-measure results better for 13 datasets out of 14. The Fig. 3 presents the AUC and F-Measure results in the graph representation.

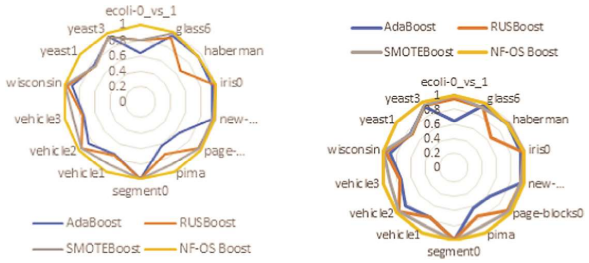

Fig. 3. AUC and F-Measure graph for NF-OS with Boosting

Hence, this section discussed several simulation results with several parameters like AUC, F-Measure, etc., and provides efficient and scalable results for class imbalance problems. Now next section will conclude this work with some future enhancements in brief.

5 Conclusion

Due to generating a lot of data virtually or online, balancing this huge data or analysing this data have raised several problems. In literature, we have seen that no major work have been done with respect to/overcome this problem. Hence, this work presents a novel approach for removing noise from the datasets using Noise Filtering (NF) and also to deal with an imbalanced classification problem by performing SMOTE after NF. Hence in this work, before training a classifier, NF first filter the noisy samples from the original dataset using K-NN technique, and then use the new minority and majority dataset to train a classifier. The experimental results over 14 datasets shows outperform of NF-OS with Boosting on AUC and F-measure. Also, this work performs the comparison among AdaBoost, RUSBoost and SMOTEBoost and produces the best results. Hence for future work, we can extend this/our work with several real world problems like balancing the Facebook data/twitter data, etc. So all the researchers, who are working in/related to this problem/area are kindly invited to do their research (in this area).

References

1. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 475–482. Springer (2009)
2. Cano, A., Zafra, A., Ventura, S.: Weighted data gravitation classification for standard and imbalanced data. IEEE Trans. Cybern. **43**(6), 1672–1687 (2013)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)

4. Codetta-Raiteri, D., Portinale, L.: Dynamic bayesian networks for fault detection, identification, and recovery in autonomous spacecraft. IEEE Trans. Syst. Man Cybern. Syst. **45**(1), 13–24 (2015)
5. Elkan, C.: The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence, vol. 17, pp. 973–978. Lawrence Erlbaum Associates Ltd. (2001)
6. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**(4), 463–484 (2012)
7. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing, pp. 878–887. Springer (2005)
8. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks, IJCNN 2008, IEEE World Congress on Computational Intelligence, pp. 1322–1328. IEEE (2008)
9. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **9**, 1263–1284 (2008)
10. Kang, Q., Huang, B., Zhou, M.: Dynamic behavior of artificial Hodgkin-Huxley neuron model subject to additive noise. IEEE Trans. Cybern. **46**(9), 2083–2093 (2016)
11. Kang, Q., Zhou, M., An, J., Wu, Q.: Swarm intelligence approaches to optimal power flow problem with distributed generator failures in power networks. IEEE Trans. Autom. Sci. Eng. **10**(2), 343–353 (2013)
12. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **39**(2), 539–550 (2009)
13. Liu, X.Y., Zhou, Z.H.: The influence of class imbalance on cost-sensitive learning: an empirical study. In: Proceedings of the Sixth International Conference on Data Mining, pp. 970–974. IEEE (2006)
14. Maciejewski, T., Stefanowski, J.: Local neighbourhood extension of smote for mining imbalanced data. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 104–111. IEEE (2011)
15. Oliker, N., Ostfeld, A.: A coupled classification-evolutionary optimization model for contamination event detection in water distribution systems. Water Res. **51**, 234–245 (2014)
16. Sáez, J.A., Galar, M., Luengo, J., Herrera, F.: INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control. Inf. Fusion **27**, 19–32 (2016)
17. Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering. Inf. Sci. **291**, 184–203 (2015)
18. Somasundaram, A., Reddy, U.S.: Modelling a stable classifier for handling large scale data with noise and imbalance. In: 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1–6. IEEE (2017)
19. Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. **40**(12), 3358–3378 (2007)
20. Tang, Y., Zhang, Y.Q., Chawla, N.V., Krasser, S.: SVMs modeling for highly imbalanced classification. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **39**(1), 281–288 (2009)

21. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7), 1088–1099 (2006)
22. Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: A novel noise filtering algorithm for imbalanced data. In: 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 9–14. IEEE (2010)
23. Yin, H.L., Leong, T.Y.: A model driven approach to imbalanced data sampling in medical decision making. In: MedInfo, pp. 856–860 (2010)
24. Zhang, Y., Zhou, Z.H.: Cost-sensitive face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1758–1769 (2010)