The Impact of Pre-processing on Classification of Imbalanced Data

Uma Salunkhe[1], Suresh Mali[2]

[1]*Department of I.T., Sinhgad College of Engineering, Pune, Maharashtra, India*

[2]*Sinhagad Institute of Technology & Science, Pune University, Pune, Maharashtra, India*

[1]urgodase.scoe@sinhgad.edu

[2]snmali_sits@sinhgad.edu

Abstract— **Machine learning is considered as a subfield of Artificial Intelligence and is concerned with the development of techniques that enable the computer to learn. One of the most common tasks in machine learning is classification which is used to organize and categorize the data so that it can be used in an efficient manner. Many real life applications suffer from imbalanced distribution of data where one class has very high number of samples relative to other class. Traditional classification algorithms may not perform efficiently in classifying such imbalanced data. Hence recent studies focus on handling imbalanced data in various ways. This paper focuses on different data level approaches used to handle imbalanced data sets by pre-processing the imbalanced data in order to convert it into balanced form. Experimental results also show the extent to which behaviour of classification algorithm is affected due to pre-processing technique.**

Keywords— **Imbalanced data, pre-processing, data level, over-sampling**

I. INTRODUCTION

Classification is the common task in machine learning that facilitates to organize and categorize data into distinct classes with the help of a model trained on the training data. Recently researchers are paying much attention to classifiers because of their applications in variety of fields like image classification, credit scoring and image retrieval. Many real life applications are suffering from imbalanced distribution of data i.e. one class has very small number of samples compared with other class. Hence, classification of imbalanced dataset has gained a great deal of attention in the past few years. Previous studies show that traditional classifiers may not give the effective performance for imbalanced dataset. This necessitates need to put efforts in modifying the existing techniques so that they can handle imbalanced data in efficient manner.

Existing literature presents four ways of resolving imbalanced dataset classification problem:

• Data level approaches: These approaches use pre-processing techniques that can convert imbalanced dataset into balanced dataset. Techniques like under sampling, oversampling or hybrid approach are the representatives of this category.

• Algorithm level approaches: In this approach existing algorithms are modified so that they can handle imbalanced distribution of data.

• Cost sensitive learning approaches: This approach is a combination of data level & algorithm level approach in which different misclassification costs are assigned to majority & minority class.

• Classifier Ensemble techniques: These techniques use ensemble of classifiers that can improve the performance of classification by combining a set of different classifiers.

Pre-processing techniques applied in data level approaches are independent of the classifier which is being used. Also they help to enhance the classification performance. In this paper, we review commonly used data level approaches in the area of imbalanced data set classification. This paper aims to explore the extent to which behaviour of pre-processing technique is affected.

The remainder of this paper is organized as follows: In Section II, we review the significant works in this area. Section III discusses the different pre-processing techniques to handle imbalanced data. It also briefs various challenges faced in the classification of imbalanced data set. Section IV reviews evaluation parameters and experimental setup. Finally last section discusses the experimental results and some findings based on those results.

II. RELATED WORK

This section briefs and categorizes the recent work carried out in the area of imbalanced data sets classification.

Yubin Park et. al. [1] proposed two types of decision tree ensembles in order to handle imbalanced class distribution. They have introduced a new splitting criterion with the help of diversification factor alpha (α). They have formed an ensemble of α tree which improves performance of imbalanced data sets in terms of AUROC. Chun-Hao Chen et. al. [2] proposed a framework that improves accuracy of existing system. Pengyi Yang et. al. [3] presented a data sampling technique, called sample subset optimization (SSO) that is able to handle different issues such as class imbalance,

small sample size and noisy data. Xu-Ying Liu et. al. [4] introduced two algorithms known as EasyEnsemble and BalanceCascade and proved improvement in terms of Area under the ROC Curve (AUC), F-measure, and G-mean values. But it does not focus on lack of comprehensibility.

Jerzy Blaszczynski et. al. [5] gave a new framework called as IIvotes that is generated by integrating SPIDER method of selective preprocessing with Ivotes ensemble. Evaluation is done in terms of Geometric Mean. Peng Shengguo Hu et. al. [6] extended Synthetic Minority Over-sampling Technique (SMOTE) in order to improve the prediction performance of minority class. Putthiporn Thanathamathee et. al. [7] used the idea that position of the separating function is dependent on the data which is at the boundary of the cluster. This idea has been combined with bootstrapping technique and evaluated in terms of average F-measure and AUC. Gregory Ditzler et. al. [8] proposed two ensemble based approaches known as Learn++.NIE and Learn++.CDS that can handle classification of imbalanced data in non stationary environment. Learn++.CDS rebalance the class distribution using Synthetic Minority class Oversampling Technique (SMOTE).

Jin Xiao et. al. [9] combined a dynamic classifier ensemble approach (DCEID) with cost sensitive learning to improve the results. DCEID combines dynamic classifier selection (DCS) and dynamic ensemble selection (DES) which are two types of dynamic classifier ensembles.

A. I. Marqués et. al. [10] formed two level composite ensembles by introducing diversity with the help of data level and feature level method. There is further scope to extend this idea and design a multilevel ensemble that uses number of ensembles jointly.

Antonio Maratea et. al.[11] presented a new accuracy measure, called as Adjusted F-measure (AGF), that takes into account the different misclassification costs of the two classes.

III. PRE-PROCESSING TECHNIQUES

a. Under-sampling

This approach applies under-sampling to the majority class by removing majority class samples from the original data. Number of instances to be removed depends on the imbalance ratio of imbalanced data set and amount of under-sampling required. The technique aims to remove redundant, noisy samples but sometimes may remove some important instances of majority class.

b. Over-sampling

This is a re-sampling approach that over samples the minority class by adding the minority samples to the original data set. If instances are added by replicating the original instances in the minority class then it may cause very specific decision regions resulting in over fitting. This problem can be resolved with the help of synthetic data generation techniques. Synthetic Minority Oversampling technique is commonly used representative of this category.

c. Hybrid

Some proposals suggest combined use of under-sampling and over-sampling which may be more effective than individual techniques.

IV. EXPERIMENTAL SETUP

The experiments were carried out using Weka environment with its default parameters. Weka is an open source toolkit that provides a set of machine learning & pre-processing algorithms.

In this study, we have implemented C4.5 classification algorithm with and without pre-processing. Experiments are carried out with Synthetic Minority Over-sampling as pre-processing technique. All experiments were carried out using 10 fold cross validation. Results of classification with pre-processing and without pre-processing are compared in order to show the extent to which pre-processing of imbalanced data affects the classification performance.

Experimental datasets

For experimentation, we have considered five imbalanced data sets that are publicly available in KEEL repository. Details of those datasets are given in table 1.

Table 1 Imbalanced Data sets

Data set	# Attributes	# Examples	# Majority	# Minority
Abalone19	9	4174	4142	32
Yeast6	9	1484	1449	35
Car-good	7	1728	1659	69
Flare-f	12	1066	1023	43
lymphography	19	148	142	6

Evaluation criteria

In order to evaluate classification systems, different parameters such as accuracy, G-mean, type-I error, type-II error, AUC (Area under ROC curve) are used. These parameters can be derived by using confusion matrix. In this paper, we have used accuracy and AUC as evaluation parameters [27].

Table 2 Confusion Matrix for a two class problem

	Predicted as Positive Prediction	Predicted as Negative Prediction
Positive Class	True Positive	False Negative
Negative Class	False Positive	True Negative

Accuracy:

It represents the proportion of the correctly classified cases on a particular data set [12]. However, accuracy does not represent the error on individual class separately as false positive and false negative have different misclassifications costs.

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Area under ROC curve(AUC):

The AUC is defined as arithmetic average of the mean predictions for each class. Previous studies suggest that area under the ROC curve (AUC) is an appropriate parameter for evaluation of imbalanced data set.

$$\text{AUC} = \frac{Sensitivity + specificity}{2}$$

$$\text{sensitivity} = \frac{\text{TP}}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

V. RESULTS AND DISCUSSION

Table 3 shows accuracy, true positive and false negative values for different imbalanced data sets classified with C4.5 classification algorithm. Results show that classifier performs very well in terms of classification accuracy. However detailed analysis in terms of true positive and false negative shows that accuracy is biased towards majority class. Consider abalone19 data set with 99.23 % accuracy. True positive value for the same data set is 0 which indicates that no element of positive (minority) class is classified correct. Value of false negative is 32 i.e. all positive elements are incorrectly classified as negative. This proves that accuracy does not reflect the errors on individual class separately.

Table 3 Accuracy of C4.5 classifier

	# Positive Instances	# Negative Instances	Accuracy	True Positive	False Negative
Abalone19	32	4142	99.23	0	32
Yeast6	35	1449	98.05	19	16
Car-good	69	1659	96.01	0	69
Flare-f	43	1023	95.97	0	43
Lymphography	6	142	97.97	3	3

Table 4 compares the AUC values of C4.5 classification algorithm with and without pre-processing. Results show that Synthetic Minority Oversampling pre-processing technique on the imbalanced training data set helps to improve the results. Though performance improvement after pre-processing is small, it is beneficial for imbalanced data set where accuracy of minority class is extremely important.

Table 4 Comparison of AUC with and without pre-processing

	No Pre-processing	SMOTE
Abalone19	0.48	0.51
Yeast6	0.81	0.81
Car-good	0.49	0.99
Flare-f	0.53	0.69
Lymphography	0.55	0.83

Fig. 1 plots AUC values of some sample datasets tested with C4.5 classification algorithm.

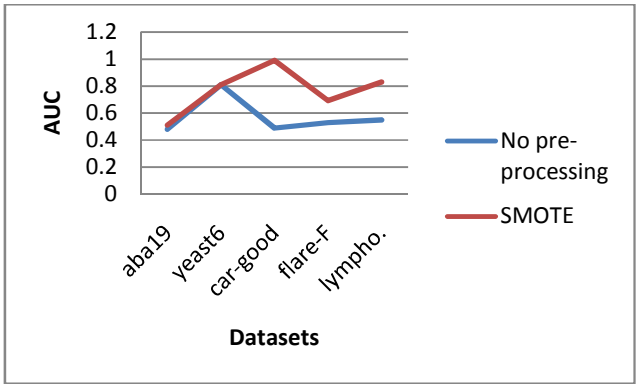

Fig. 1 comparison of performance with and without pre-processing

Results show that pre-processing of imbalanced data improves the performance in terms of AUC. Though same over-sampling procedure is applied on all the datasets, they have shown variations in their improvement. This is due to other challenges raised by imbalanced nature of data. This necessitates need to focus on other factors associated with imbalanced distribution.

VI. CONCLUSIONS

In this paper, we have presented a survey of different data level approaches used to handle imbalanced data sets. We also assessed the evaluation parameters for classification and

results show that accuracy does not represent different misclassification costs of different classes. Comparison of the results of classification with and without pre-processing shows that pre-processing improves the performance of classification. It can be further enhanced by overcoming limitations of existing pre-processing techniques.

REFERENCES

[1] Yubin Park and Ghosh, J., "Ensembles of α-Trees for Imbalanced Classification Problems,", IEEE Transactions on Knowledge and Data Engineering, vol.26, no.1, pp.131-143, January 2014.

[2] Chun-Hao Chen et. al., "A combined mining-based framework for predicting telecommunications customer payment behaviors," Expert Systems with Applications, vol.40, no.16, pp.6561 - 6569, November 2013.

[3] Pengyi Yang et. al., "Sample Subset Optimization Techniques for Imbalanced and Ensemble Learning Problems in Bioinformatics Applications," IEEE Transactions on Cybernetics, vol .44, no. 3, pp. 445 - 455, March 2014.

[4] Xu-Ying Liu, Jianxin Wu and Zhi-Hua Zhou, "Exploratory Under-sampling for Class-Imbalance Learning, " IEEE Transactions on Systems, Man, and Cybernetics, vol.39, no.2, pp.539 - 550, April 2009.

[5] Jerzy B laszczynski, Magdalena Deckert, Jerzy Stefanowski, SzymonWilk, "Integrating Selective Pre-processing of Imbalanced Data with Ivotes Ensemble," Knowledge Engineering , vol. 21, no. 9, pp. 1263–1274, 2010.

[6] Shengguo Hu, Yanfeng Liang, Lintao Ma, Ying He, "MSMOTE: Improving Classification Performance When Training Data is Imbalanced," Second International Workshop on Computer Science and Engineering , vol.2, pp.13-17, October 2009.

[7] Putthiporn Thanathamathee and Chidchanok Lursinsap, "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques," Pattern Recognition Letters, vol. 34, no. 12, , pp. 1339-1347, September 2013.

[8] Gregory Ditzler and Robi Polikar , "Incremental Learning of Concept Drift from Streaming Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 10, pp. 2283 – 2301, October 2013.

[9] Jin Xiao, Ling Xie, Changzheng He and Xiaoyi Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," Expert Systems with Applications, vol. 39,no. 3, pp. 3668-3675, February 2012.

[10] A.I. Marqués, V. García and J.S. Sánchez, "Two-level classifier ensembles for credit risk assessment ," Expert Systems with Applications, vol. 39, no. 12, pp. 10916-10922, September 2012.

[11] Antonio Maratea, Alfredo Petrosino and Mario Manzo, "Adjusted F-measure and kernel scaling for imbalanced data learning," Information Sciences, vol. 257, pp.331 - 341, February 2014.

[12] Mikel Galar, Alberto Fernández, Edurne Barrenechea and Francisco Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary under-sampling," Pattern Recognition, vol. 46, no. 12,pp. 3460-3471, December 2013.

[13] Peng Cao et. al., "Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD," Computerized Medical Imaging and Graphics, vol. 38, no. 3, pp. 137–150, April 2014

[14] Nitesh V. Chawla, Bowyer, Hall, Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, June 2002

[15] Andrea Dal Pozzolo et. al. ," Learned lessons in credit card fraud detection from a practitioner perspective," Expert Systems with Applications, vol. 41, no. 10, pp. 4915 - 4928, August 2014.

[16] Bartosz Krawczyk, Michał Woźniak and Gerald Schaefer," Cost-sensitive decision tree ensembles for effective imbalanced classification," Applied Soft Computing, vol. 14, pp. 554-562, January 2014.

[17] Maciej Zięba, Jakub M. Tomczak, Marek Lubicz and Jerzy Świątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," Applied Soft Computing, vol. 14, pp. 99-108, January 2014.

[18] Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong and Yang Wang, "Cost-sensitive boosting for classification of imbalanced data," Pattern Recognition, vol. 40, no. 12, pp. 3358-3378, December 2007.

[19] Bhowan, U., Johnston, M., Mengjie Zhang and Xin Yao, "Evolving Diverse Ensembles Using Genetic Programming for Classification With Unbalanced Data," IEEE Transactions on Evolutionary Computation, vol. 17, no. 3, pp. 368-386, June 2013.

[20] Castro, C.L. and Braga, A.P., "Novel Cost-Sensitive Approach to Improve the Multilayer Perceptron Performance on Imbalanced Data," IEEE Transactions on Neural Networks and Learning Systems, vol. 24, no. 6, pp. 888 - 899, June 2013.

[21] Matías Di Martino et. al., "Novel classifier scheme for imbalanced problems," Pattern Recognition Letters, vol. 34, no. 10, pp. 1146 - 1151, July 2013.

[22] Joaquin Abellan and Carlos Mantas, "Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring", Expert Systems with Applications, vol. 41,no. 8, pp. 3825-3830, June 2014.

[23] Gang Wang, Jian Ma and Shanlin Yang, "An improved boosting based on feature selection for corporate bankruptcy prediction," Expert Systems with Applications, vol. 41, no. 5, pp. 2353-2361, April 2014.

[24] Lu Han, Liyan Han and Hongwei Zhao,"Orthogonal support vector machine for credit scoring", Engineering Applications of Artificial Intelligence, , vol. 26, no. 2, pp. 848–862, February 2013.

[25] A.I. Marqués, V. García and J.S. Sánchez , "Exploring the behavior of base classifiers in credit scoring ensembles", Expert Systems with Applications, vol. 39 , no. 11 , pp. 10244–10250, September 2012.

[26] Galar et. al., "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, vol. 42, no. 4, pp. 463-484, July 2012.

[27] Salvador Garcıa et. al. ," Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems," Applied Soft Computing, vol. 9, no. 4, pp. 1304–1314 , May 2009.

Book

[1] Ludmila I. Kuncheva ,Combining Pattern Classifiers Methods and Algorithms, A John Wiley & Sons Inc. Publication, 2004.

[2] Christopher M. Bishop ,Pattern Recognition and Machine Learning, Springer, 2006.