Cost Sensitive Learning and SMOTE Methods for Imbalanced Data

Akbar Khan[1], Faizullah Khan[2], Surat Khan[3],Ishtiaq Ahmed Khan[3], Muhammad Saeed[4]

[1]Department of Computer Engineering, [2]Department of Telecom Engineering, [3]Department of Electrical Engineering, Faculty of Information and Communication Technology, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan, [4]Department of Electrical Engineering, National University of Computer and Emerging Sciences Islamabad, Pakistan

Abstract

Class imbalance is one of the main problem using different algorithms used in machine learning. In imbalance classification of data the false negative is always high. The researchers have introduced many methods to deal with this problem, but the purpose of this paper is to apply machine learning algorithms under the SMOTE and cost sensitive learning approaches and acquired the results from the different experiments to find out the suitable results for imbalanced data.

Keywords: Cost Sensitive Learning; Machine Learning; WEKA; SMOTE; Imbalanced Data

Corresponding author's email: *Akbar.khan@buitms.edu.pk*

INTRODUCTION

In binary classification, class imbalance can be elaborated as majority class outnumbering of the minority class. This can be seen in many machine learning and data mining application such as fraud detection smoke detection (people ought to smoking) and many more (Nitesh et al., 2004). Class imbalance is also considered as ten challenging problem in data mining. Researcher, have brought and introduced different methods in order to see class imbalance (Qiang et al., 2006). Cost sensitive learning has also used while focusing imbalance using DT, NB, and more (Sheng et al., 2005; Xiaoyong et al., 2004). In a binary classification problem, majority under sampling removes instances from the majority i.e. larger class, with the aim of improving bias of the minority class instances. The majority under sampling techniques include random under sampling, Wilson's editing, and one-sided selection. Random under sampling is an effective technique in which a portion of the majority class instances are removed at random from the dataset (Jason et al., 2007; Gary et al., 2004; Gary et al., 2003).

MATERIALS AND METHODS

Data Structure

The general structure for info information with regards to data mining calculations comprise of a table like structure where information is put away in lines and segments. Each line comprises of information identified with each other, separated up in various segments where every section can have its own information sort The classification of target class contains two means binary class.

- Patient has trachoma
- Patient has no trachoma means normal

This generates a problem that is known as binary class.

Data source and Data collection

The primary data was collected using a questioners based which includes a questions related to several personal, socio economic, psychological and behavioral factors. The key looking at unit was finished Cluster/Villages. The amount of gatherings per Evaluation Unit (EU) for each association has been determined. A multi-stage pack examining system - to make sense of which gatherings and family units will be investigated in the midst of the study. With the help of an eye expert and their team has examined each and every man and woman individually and concluded their case result on the basis of affected condition

Data preprocessing

The first step of data preprocessing is the the removal of unnecessary questions from the given questioners and to elaborate of all those attributes which are most important. Every attribute holds categorical values and has represented individually in WEKA environment some of them are shown as below.

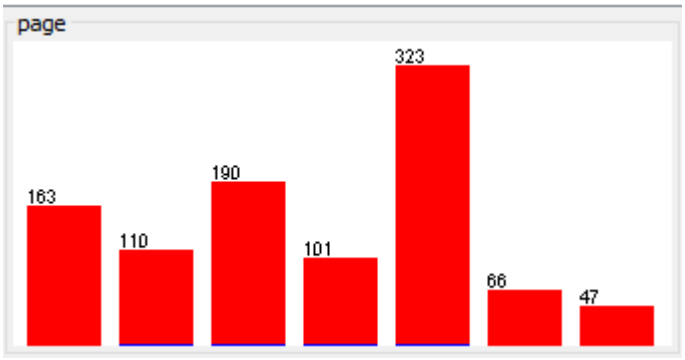

Figure 1: In the given snippet person age has shown in a categorical order using WEKA

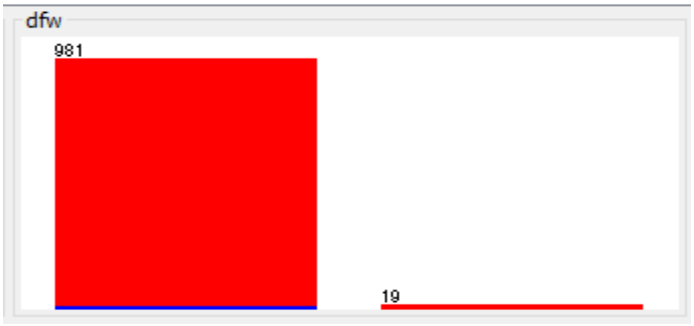

Figure 2: Dfw stands for daily face washing having the categorical value “YES” or “NO” in which there are 981 who said yes and rest of the No

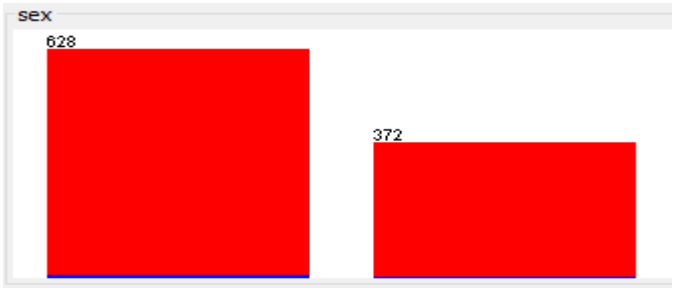

Figure 3: This graph shows SEX in which there were 628 males and 372 were female

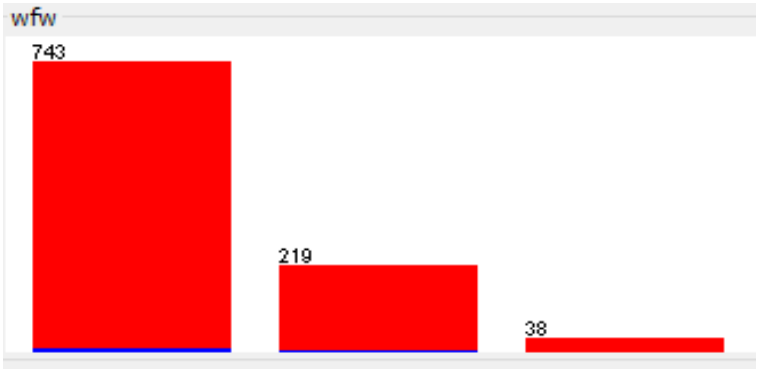

Figure 4: Wfw water for face washing in which 743 persons obtained by well, 219 piped and 38 obtained water through ground

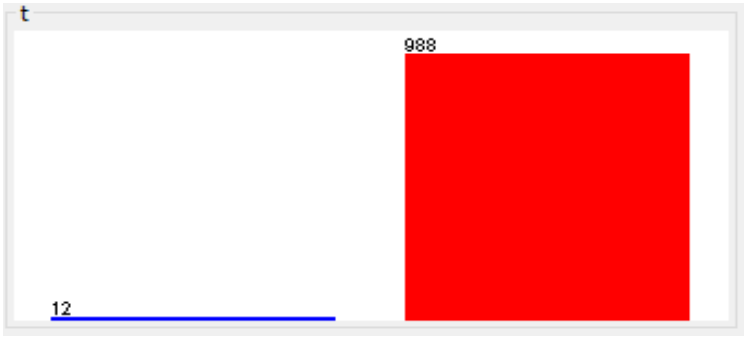

Figure 5: Trachoma T stands for trachoma the target class having categorical vale in which 12 were positive and 988 were negative

RESULTS AND DISCUSSION

The basic concept of cost sensitive learning is cost matrix. Cost-sensitive learning makes use of associating appropriate cost with incorrect classifications. It represents the penalty imposed on incorrect classification in numerical form. In binary the class can be c (min, maj). The major aim of cost sensitive learning method is to have a hypothesis that reduces the cost on the data (overall). It is beneficial for sampling methods therefore it provides a feasible alternative to sampling methods for imbalanced data. It attempts to balance the distribution (Elkan, 2001).

Results of cost sensitive classifier

Table 1: Complete results of cost sensitive classifier

Confusion matrix	Sensitivity	Specificity	AUC	Algorithms
0 1 1 0	0	1	0.5	SVM
	0	0.97	0.394	LR
	0	1	0.433	J48
	0	1	0.568	RF
0 2 1 0	0	1	0.5	SVM
	0	0.994	0.397	LR
	0	1	0.433	J48
	0	0.998	0.562	RF
0 5 1 0	0	1	0.5	SVM
	0.985	0.083	0.393	LR
	0	0.995	0.401	J48
	0	0.994	0.556	RF
0 10 1 0	0	1	0.5	SVM
	0.083	0.982	0.407	LR
	0	0.986	0.454	J48
	0	0.987	0.548	RF
0 15 1 0	0	0.997	0.499	SVM
	0.166	0.972	0.373	LR
	0	0.960	0.383	J48
	0	0.960	0.545	RF
0 20 1 0	0	0.980	0.49	SVM
	0.166	0.962	0.408	LR
	0	0.951	0.43	J48
	0	0.949	0.543	RF

Confusion matrix	Sensitivity	Specificity	AUC	Algorithms
2 1 3 0	0	1	0.5	SVM
	0	0.97	0.394	LR
	0	1	0.433	J48
	0	1	0.568	RF
2 2 3 0	0	1	0.5	SVM
	0	0.994	0.397	LR
	0	1	0.433	J48
	0	0.998	0.562	RF
2 5 3 0	0	1	0.5	SVM
	0.985	0.083	0.393	LR
	0	0.995	0.401	J48
	0	0.994	0.556	RF
2 10 3 0	0	1	0.5	SVM
	0.083	0.982	0.407	LR
	0	0.986	0.454	J48
	0	0.987	0.548	RF
2 15 3 0	0	0.997	0.499	SVM
	0.166	0.972	0.373	LR
	0	0.960	0.383	J48
	0	0.960	0.545	RF
2 20 3 0	0	0.980	0.49	SVM
	0.166	0.962	0.408	LR
	0	0.951	0.43	J48
	0	0.949	0.543	RF

SMOTE (Synthetic Minority Oversampling Technique)

SMOTE (synthetic minority oversampling technique) is a powerful method which is usually used in imbalance data when the one class is less than other class. When there is binary classification of class means positive and negative so in this regard SMOTE is a special technique which is widely used in this type of situation. There are mainly two techniques in SMOTE namely oversampling and under sampling. The concept of oversampling is to increase the number of minority class where the concept of under sampling is to minimize the number of negative class. It is widely used in imbalance data (Haibo et al., 2009).

Table 2: Complete result of SMOTE

Sensitivity	Specificity	AUC	Algorithms
0	1	0.5	SVM
0	0.97	0.394	LR
0	1	0.433	J48
0	1	0.568	RF
0	1	0.5	SVM
0	0.996	0.699	LR
0	1	0.449	J48
0	0.998	0.745	RF
0	1	0.5	SVM
0	0.997	0.811	LR
0	1	0.482	J48
0	0.998	0.841	RF
0	1	0.5	SVM
0.510	0.968	0.867	LR
0.229	0.979	0.699	J48
0.510	0.969	0.896	RF

CONCLUSION

The original data was picked up from the expert in the domain from Helper Eye Hospital. In which they have checked more than 1000 patients and found some patients causing Trachoma. The data was too imbalance with 12 positive and 988 was negative applied under sampling and oversampling on the data, and it has

been calculated from the above mentioned tables for SMOTE and Cost Sensitive Learning, that the results of SMOTE is accurate in the sense of good AUC. SMOTE performs better than cost sensitive classifier for 96+ve instances and 988 negative instances.

REFERENCES

- Nitesh VC, Nathalie J, Aleksander K. (2004). Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1):1–6.
- Qiang Y, Xindong W. (2006). 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making 5(4):597–604.
- Sheng S, Ling CX, Yang Q. (2005). Simple test strategies for cost sensitive decision trees. Springer-Verlag Berlin Heidelberg, pp. 365–376.
- Xiaoyong C, Lin D, Qiang Y, Charles X Ling. (2004). Test-cost sensitive naïve Bayes classification, in International Conference on Data Mining, pp. 51–58.
- Jason VH, Taghi M K, Amri N. (2007). Experimental perspectives on learning from imbalanced data, In Proceedings of the 24th International Conference on Machine Learning, Corvallis, pp.935-942.
- Gary MW. (2004). Mining with rarity: a unifying framework. ACM SIGKDD Explorations 6(1):7-19.
- Gary MW, Foster P. (2003). Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research 19:315-354.
- Elkan C. (2001). The Foundations of Cost-Sensitive Learning. Proc. Int'l Joint Conf. Artificial Intelligence, pp. 973-978.
- Haibo H, Edwardo G. (2009). Learning from imbalanced data, IEEE Trans. Knowl. Data, pp 1263-1284.