SMOTE bagging algorithm for imbalanced dataset in logistic regression analysis (case: Credit of bank X)

Article · January 2015

DOI: 10.12988/ams.2015.58562

3 authors, including:

Hari Wijayanto
Bogor Agricultural University
46 PUBLICATIONS **110** CITATIONS

SEE PROFILE

Anang Kurnia
IPB University (Bogor Agricultural University)
90 PUBLICATIONS **134** CITATIONS

SEE PROFILE

Applied Mathematical Sciences, Vol. 9, 2015, no. 138, 6857 - 6865
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.58562

SMOTE Bagging Algorithm for Imbalanced Dataset in Logistic Regression Analysis (Case: Credit of Bank X)

Fithria Siti Hanifah

Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Hari Wijayanto

Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Anang Kurnia

Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Abstract

Logistic regression analysis is one of classification methods which is both most popular and common used. This classifier works well when the class distribution in response variable is balanced. In many real cases, the imbalanced class dataset frequently was found. This problem can affect of being difficult at obtaining a good predictive model for minority class dataset. The prediction accuracy generated will be good for majority class but not for minority class. SMOTEBagging is a combination of SMOTE and Bagging algorithm which is used to solve this problem. The purpose of this study is to create a powerful model at classifying the imbalanced data and to improve the classification performance of weak classifier. This study used credit scoring data which is imbalanced data consisting of 17 explanatory variables involved. The result from this study showed that the sensitivity and AUC value from SMOTEBagging Logistic Regression

(SBLR) model is greater than the sensitivity and AUC value of logistic regression model. Moreover, SMOTEBagging algorithm can increase the accuracy of minority class.

Keywords: Accuracy, Imbalanced data, Logistic regression analysis, SMOTEBagging

1. Introduction

In many cases of the classification, the common problem is the imbalanced data. Imbalanced dataset occurs when there are one or more classes that dominate the overall dataset as a majority class and the other class which is a rare occurrence as minority class. The problem of imbalanced dataset occurs in many cases of classification, such as the classification of poverty [8], text classification [3], the classification of the success of a student's study [9], medical diagnosis [11], credit scoring [2], etc.

One of classification methods that is popular and often be used is logistic regression. This method works well in classifying when the class distribution of response variables in the dataset is balanced. However, if the dataset used is imbalanced, it will have an impact on the difficulty of getting a good predictive and meaningful model due to the lack of information from the minority class [11]. This standard method will produce a bias toward the classes with a greater number of instances (the majority) because the classifier will tend to predict the majority class data. The minority class will be ignored (treating them as a noise), so the observation from the minority class cannot be classified correctly [5].

There are three approach into three groups, there are the algorithm level, the data level, and the cost-sensitive which is a combination of algorithm level and the data level [5]. Chawla *et al.* [3] introduced SMOTE is a method that was developed based on the concept of oversampling. SMOTE works by generating synthetic samples based k-nearest neighbors. It was expected to handle the weaknesses of undersampling method based that eliminates the important information in the data.

In addition, the ensemble methods can also be used for imbalanced data problem by improving the accuracy of single classifier. This methods combining decisions of several different classifier to output a single class label by voting process.

This study will apply the SMOTEBagging method on the classification of logistic regression model. SMOTEBagging is a combination of SMOTE and ensemble Bagging algorithm, where the SMOTE will be involved in the process of Bagging, generating synthetic samples on data subset from Bootstrap. The data used is the credit scoring from one of bank in Indonesia.

2. Logistic Regression

Logistic regression model is a modeling procedure applied to model the response variable Y that is category based on one or more of the predictor variables X, whether it is a category or continuous [1]. The binary logistic regression model can be formulated as follows:

$$E(y \mid x) = \pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}$$

Equivalently, the log odds which is also called the logit, has the linear formula as follows:

$$\text{logit}[\pi_i] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Hosmer and Leme show [7] said that the common fathoming of logistic regression parameter is a maximum probability method. The principle of this method is to maximize the probability of functions.

Parameters significance testing simultaneously performed with the Probability Ratio Test. The Probability Ratio Test can be formulated as follows:

$$G = -2\log\left(\frac{l_0}{l_1}\right) = -2\left[\log(l_0) - \log(l_1)\right]$$

Where l_0 is the maximum value of the probability function under Ho, l_1 is the maximum value of the probability function under alternative Ha.

The G statistic follows the chi-square distribution with $df = p$. If using the real level of α, the test criteria is rejected H_0 if $G \geq \chi^2_{(p)}$ or $p - value \leq \alpha$, which means received in other cases [1].

Partial parameters significance testing is performed by the Wald Test. The Wald Test can be formulated as follows:

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

$SE(\hat{\beta}_i) \equiv$ *Standard eror* from $\hat{\beta}_i$.

3. SMOTE Bagging

SMOTEBagging is a combination of SMOTE and Bagging algorithm. SMOTEBagging involves generation step of synthetic instances during subset construction. [10]. *Synthetic Minority Oversampling Technique* (SMOTE) is one of the oversampling methods that has been first introduced by Chawla *et al.* [3]. SMOTE works by generating synthetic data as far as the amount of minority data equivalent to the majority data. Synthetic data is made based on the characteristic of the object and k-nearest neighbor. Bagging is an abbreviation of Bootstrap Aggregatting introduced by Breiman (1996) with the purpose to reduce the variance

of predictors. Zhou [12] stated that the basic idea of ensemble method is bootstrap, to generate a new dataset to create a classifier in many versions. The purpose of this combination is to create a powerful model in classifying imbalanced data without sacrificing overall accuracy.

According to SMOTEBagging, each subset is obtained from the Bootstrap process balanced by SMOTE before the modeling. Two parameters need to be decided in SMOTE: k-nearest neighbors and the total number of over-sampling from minority class – N. The total of over-sampling decided as far as amount of majority class and minority class is balance.

4. Performance Measurements

Evaluation of a classification algorithm performance is measured by confusion matrix. Confusion matrix contains information about the actual and the prediction class presented in the following table:

Table 1. The Confusion Matrix

Actual	Prediction	
	Positive class	Negative class
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

Evaluating the results of the classification based on the value of the confusion matrix measured by calculating the value of accuracy, sensitivity, and specificity. The accuracy shows the overall level of accuracy, the sensitivity shows the accuracy in class-i whereas the specificity shows the accuracy in class-j. The classification performance evaluation formula is:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

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

The accuracy of classification can also be measured by calculating area under curve (AUC) in Receiver Operating Characteristic (ROC) analysis. According to Fawcett [4] ROC curve illustrates the classification performance in two dimensions: probability plot of negative false (1-specificity) with the correct prediction of positive true (Sensitivity). AUC values ranged from 0 to 1. If the AUC values near to 1 means the model accuracy or classification is high.

5. Result and Discussion

The result of the probability ratio test (G^2) is at 595.00 and $\chi^2_{(17,0.05)} = 8.67$, so Ho is rejected. It means, with a significance level of 5%, the data showed at

least there is an explanatory variable that affect the model. Then, the testing was continued with the partial testing parameters by using the Wald test. The results which were obtained indicating several explanatory variables that significantly affected the model. In addition, the variable reduction was used by using a forward stepwise. The result of logistic regression model for training data and testing data, showed in Table 2 below:

Table 2 The Performance of Classification Logistic Regression

	Training	Testing
Accuracy	88.62%	86.00%
Sensitivity	8.16%	0.0%
Specificity	99.86%	98.85%
AUC	75.20%	77.01%

Table 2 showed the classification of logistic regression performance for the whole dataset (consolidated training and validation sample/ training and testing). The "goods" are denoted as "negative", while the "bads" are denoted as "positive" due to the predicting of the "bads" are more important in credit scoring development. AUC values for training and testing data show that the accuracy model is good enough. Presentation of model accuracy in classifying the loan customers is quite high. However, there is an imbalance in predicting credit that can be seen in a very small value in sensitivity model. This imbalance leads the prediction model directed to the majority class, namely good customer. Thus, the prospective customers who applies for loan is likely to be accepted.

The Model with SMOTEBagging

The applied SMOTEBagging algorithm was decided from several parameters, which were the number of bootstrap (10, 20, 25, 30,….,1000), k-nearest neighbor (1-10) and total number of oversampling (100, 200,….,500). The figure below illustrates the results of performance measurements for each bootstrap replication.

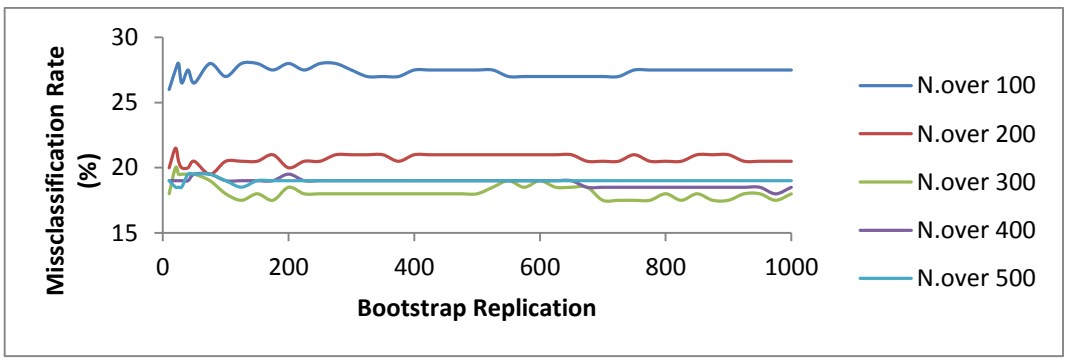

Figure 1. The Missclassification rate for each bootstrap replication

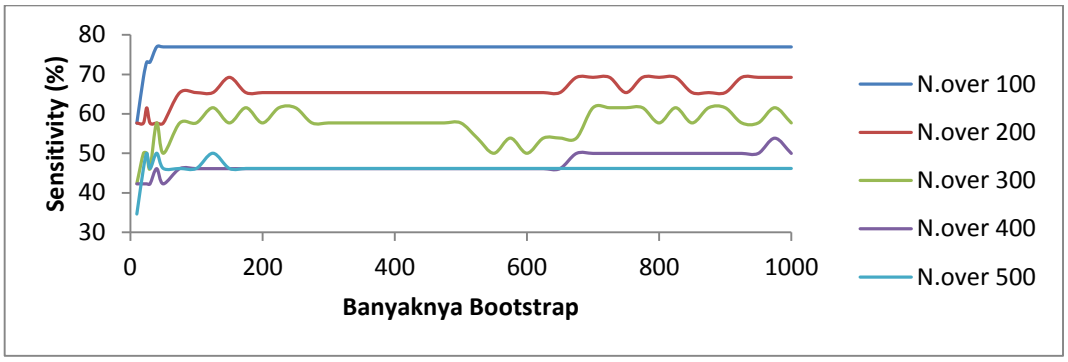

Figure 2. Sensitivity for each bootstrap replication

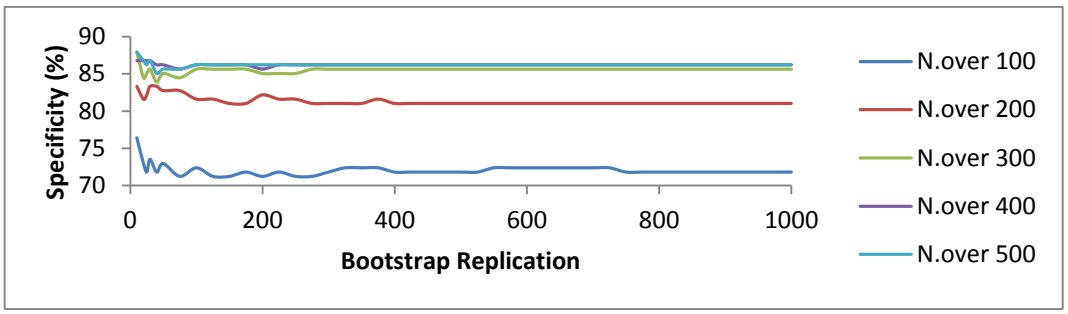

Figure 3. Sensitivity for each bootstrap replication

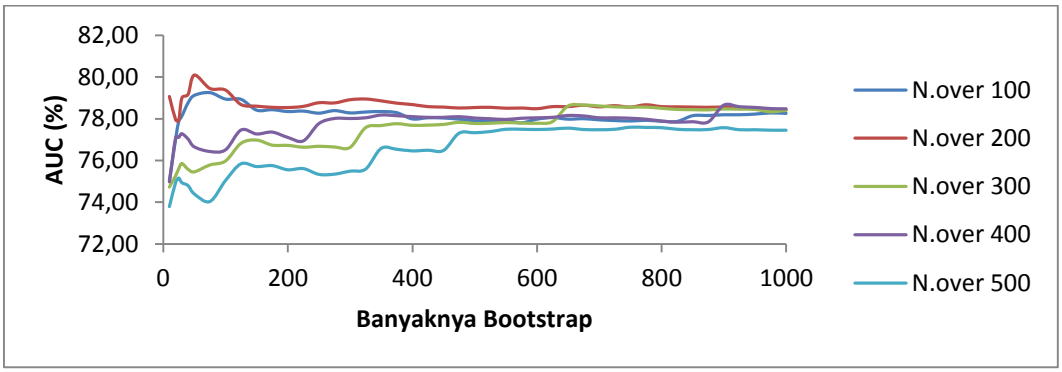

Figure 4. AUC values for each bootstrap replication

Table 3 The Performance of Classification based on k-Nearest Neighbour

Model	K	Accuracy	Sensitivity	Specificity	AUC
SMOTE Bagging Logistic Regression	1	81.00%	57.69%	84.48%	78.90%
	2	82.00%	65.38%	84.48%	79.19%
	3	80.00%	61.54%	82.76%	80.06%
	4	**80.00%**	**61.54%**	**82.76%**	**80.14%**
	5	79.50%	57.69%	82.76%	80.10%

Table 3 (Continued): The Performance of Classification based on k-Nearest Neighbour

SMOTE Bagging Logistic Regression	6	80.50%	61.54%	83.33%	78.92%
	7	79.50%	57.69%	82.76%	78.17%
	8	80.00%	65.38%	82.18%	78.39%
	9	80.00%	65.38%	82.18%	79.05%
	10	79.00%	65.38%	81.03%	79.20%

Based on the previous results, the optimal SMOTEBagging logistic regression (SBLR) model is obtained from a parameters combination, the number of bootstrap replicate as much as 50 times, N oversampling as much as 200, and k-nearest neighbour is 4.

Model Comparison

The model which has been obtained then was compared with the level of accuracy by AUC value (Table 4). AUC value in SMOTEBagging logistic regression (SBLR) model 3.13% higher than the logistic regression model. It shows in the model with SMOTEBagging is slightly more accurate than the model without it.

In addition to comparing the AUC values, we also considered the sensitivity and the specificity of each model. AUC value itself composed from sensitivity and specificity. SMOTEBagging model can improve the sensitivity quite large although the specificity slightly decreased. The sensitivity of logistic regression model is 0, it means this model predicted all data into “goods”. The sensitivity of SMOTEBagging logistic regression (SBLR) model was higher than logistic regression model, meanwhile the value of specificity logistic regression model was slightly larger.

Table 4 Perfomance comparison Logistic Regression vs SBLR

	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	86%	0.0%	98.85%	77.01%
SBLR	80.00%	61.54%	82.76%	80.14%

The ROC curves in Figure 5 showed that the horizontal scale represented a false positive rate (1-specificity) and the vertical scale represents a true positive rate (sensitivity). Based on ROC curve, the logistic regression model with SMOTEBagging is generally better than the logistic regression model.

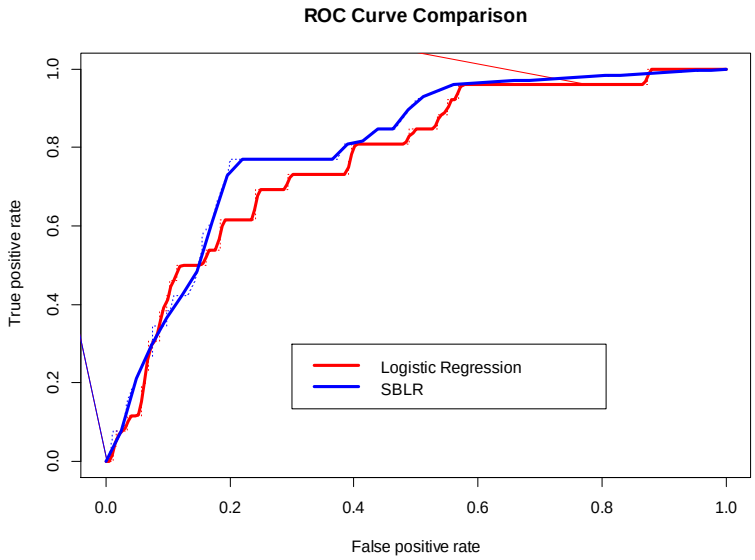

Figure 5. The ROC Model Comparison Curve

6. Conclusion

In this case, credit scoring data classification with imbalance rate at 12,4% minority and 87,6% majority, showed that SMOTEBagging logistic regression (SBLR) model is more accurate than the logistic regression model. It is shown by the AUC value generated by SMOTEBagging logistic regression (SBLR) model is higher than the logistic regression model. Moreover, the accuracy of minority class (sensitivity) in SMOTEBagging model is much better. The result indicates that SMOTEBagging can increase the level of accuracy model in imbalanced data. To assess the performance stability of this algorithm, we need simulate in various imbalance rate.

References

[1] A. Agresti, *Categorical Data Analysis*, John Willey & Sons, Inc, New York, 2002.

[2] I. Brown, and C. Mues, An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets, *Expert Systems with Applications*, **39** (2012), no. 3, 3446-3453. http://dx.doi.org/10.1016/j.eswa.2011.09.033

[3] V.N. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, *Journal of Artificial Intelligence Research*, **16** (2002), 321-357.

[4] T. Fawcett, An Introduction to ROC analysis, *Pattern Recognition Letters*, **27** (2006), 861-874. http://dx.doi.org/10.1016/j.patrec.2005.10.010

[5] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, 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)*, **42** (2011), 463-484. http://dx.doi.org/10.1109/tsmcc.2011.2161285

[6] H. He, E.A. Garcia, Learning from Imbalanced Data, *IEEE Transaction on Knowledge and Data Engineering*, **21** (2009), 1263-1284. http://dx.doi.org/10.1109/tkde.2008.239

[7] D.W. Hosmer, and S. Lemeshow, *Applied Logistic Regression*, John Willey and Sons, Inc, New York, 1989.

[8] M.J. Muttaqin, B.W. Otok, and S.P. Rahayu, *Metode Ensemble pada CART untuk Perbaikan Klasifikasi Kemiskinan di Kabupaten Jombang*, [Undergraduate Thesis], Surabaya, Sepuluh November Institut of Technology, 2013.

[9] H. Rahmah, *Penerapan Smote Pada Metode Cruise Untuk Penentuan Faktor Keberhasilan Studi Mahasiswa BUD*, [Undergraduate Thesis], Bogor, Bogor Agricultural University, 2013.

[10] S. Wang, X. Yao, Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models, *IEEE Symp. Comput. Intell. Data Mining*, (2009), 324–331. http://dx.doi.org/10.1109/cidm.2009.4938667

[11] B.W. Yap, K. Abd Rani, H.A. Abd Rahman, S. Fong, Z. Khairudin, N.N. Abdullah, An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets, *Proceedings of the First International Conference on Advanced Data and Information Engineering.* Singapore, (2014). http://dx.doi.org/10.1007/978-981-4585-18-7_2

[12] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, Florida, 2012.

Received: September 15, 2015; Published: November 25, 2015