Indonesian Journal of Electrical Engineering and Computer Science
Vol. 13, No. 1, January 2019, pp. 102~108
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i1.pp102-108

Integration of synthetic minority oversampling technique for imbalanced class

Noviyanti Santoso, Wahyu Wibowo, Hilda Himawati
Department of Business Statistics, Faculty of Vocational,
Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo-Surabaya, Indonesia

Article Info

Article history:

Received Aug 1, 2018
Revised Nov 1, 2018
Accepted Nov 19, 2018

Keywords:

Accuracy
Data mining
Imbalanced class
SMOTE

ABSTRACT

In the data mining, a class imbalance is a problematic issue to look for the solutions. It probably because machine learning is constructed by using algorithms with assuming the number of instances in each balanced class, so when using a class imbalance, it is possible that the prediction results are not appropriate. They are solutions offered to solve class imbalance issues, including oversampling, undersampling, and synthetic minority oversampling technique (SMOTE). Both oversampling and undersampling have its disadvantages, so SMOTE is an alternative to overcome it. By integrating SMOTE in the data mining classification method such as Naive Bayes, Support Vector Machine (SVM), and Random Forest (RF) is expected to improve the performance of accuracy. In this research, it was found that the data of SMOTE gave better accuracy than the original data. In addition to the three classification methods used, RF gives the highest average AUC, F-measure, and G-means score.

Corresponding Author:

Noviyanti Santoso,
Department of Business Statistics, Faculty of Vocational,
Institut Teknologi Sepuluh Nopember,
Kampus ITS Sukolilo-Surabaya, 60111, Indonesia.
Email: noviyanti_s@statistika.its.ac.id

1. INTRODUCTION

A class on a dataset with unbalanced class distribution makes classification results more likely to belong to majority class than the minority class. Class imbalance in the dataset is a problem in machine learning, where the majority (negative) class is higher than the minority (positive) class. The issue of class imbalance is a common problem found in the dataset in various fields, including bankruptcy prediction, credit card fraud detection [1], and disease diagnosis [2]. Class imbalance is very disserving for researchers that are engaged in data mining. The reason is in the data mining generally has difficulties in classifying the minority class correctly. That algorithm assumes that the tested class distribution has already balanced so that there is an error in classifying the value of each class. Moreover, machine learning algorithms are designed to generalize the tested data as equal and make the simplest hypothesis. The principle is embedded in various algorithms such as decision tree, nearest neighbor, and support vector machine. Therefore, when this algorithm tests the unbalanced dataset, it will tend to focus on a majority and ignore the minority class and causing errors in minority class classification. Minority class is considered as noise only.

The problem of classification testing method in imbalanced datasets usually have the characteristics as classified instance values (misclassification cost) in the minority class higher than the misclassification in majority class. Many of research [3-6] have provided the relevance of this matter in classification case. In recent years, classification problem for imbalanced datasets became the challenging research topic. Therefore, the challenge in overcoming this is how to classify minority class more accurately. According to research [7], the way to overcome the class imbalance is to resampling the original dataset, either in minority class (oversampling), or majority class (undersampling).

Oversampling is a mechanism for balancing class distribution by randomly replicating minority class instance. However, the lack of oversampling is the increased possibility overfitting, because this procedure makes the duplication of instances precisely. Undersampling is a procedure for balancing class distribution by randomly subtracting the majority class instance. The lack of undersampling is the loss of the essential data for the continuity of the decision making the process by machine learning [8]. Then [9] proposed a solution called Synthetic Minority Oversampling Technique (SMOTE). SMOTE can generate synthetic minority sample class utilizing interpolation processes between minority class instance that located adjacent. SMOTE utilizes the nearest neighbors factor and the desired oversampling level.

There are several works that integrated SMOTE and data mining technique. According to [10] combination among SMOTE and Tomek links as resampling approach shown better performance in imbalanced class dataset. Furthermore, conclusion of [11] is AUC score of imbalanced dataset which had resampling using SMOTE increasing as well as performance of accuracy for all data mining methods that integrated in. In the medical dataset, [12] applied SMOTE ensembled machine learning approach to predict diabetes mellitus, the results is Random Forest (RF) and Naïve Bayes shown the greater score for all evaluation measurements. While [13] and [14] conclude that SVM and C. 45 is outstanding methods to predict kind of fish based on DNA barcode. There is no exactly approach that consistency provides appropriate performance, because it depends on quality and characteristic of its dataset.

Based on the description above, this research will integrate the SMOTE and data mining classification methods of Naive Bayes, SVM, and RF to evaluate their performance to overcoming unbalanced class on banking case. The results of this study are expected to be an alternative in settlement of classification cases with unbalanced classes in various fields. So it can be an early warning model to predict the events that will come with a high degree of accuracy.

2. RESEARCH METHOD

2.1. Dataset

This research is using Bank Marketing datasets from UCI Machine Learning. From 45210 instances, as much as 10% sample is randomly taken so that the number of instances used is 4521. The total of 521 instances (13%) belong to minority (positive) class, and 4000 instances (87%) is including majority (negative) class. It indicates that the Bank Marketing dataset has an unbalanced class category.

To evaluate the model, we split dataset into training set and testing set in four combinations, i.e., 90:10, 80:20, 70:30, and 50:50. Perform the validation using 5-fold cross validation then calculate the accuracy of classification using three evaluation measures, i.e., AUC, G-means, and F-measure.

2.2. Methods

2.2.1 Synthetic Minority Oversampling Technique

The SMOTE method proposed by [9] as one of the solutions in dealing with unbalanced data with the different principle from the previously proposed oversampling method. When the oversampling has various principles randomly, SMOTE method adds the number of minor class to equal to major class by generating artificial data. The artificial or synthesis data is made based on a k-nest neighbor. Determining the number of k-nest neighbors by considering the ease of the application. Generating artificial numerical data is different from categorical data. Measuring the distance of numerical data using Euclidean distance, where categorical data is simpler than numerical data, it measured by the mode value. Generating new data, in general, using (1).

$$x_{syn} = x_i + \delta \times (x_{knn} - x_i) \tag{1}$$

2.2.2 Naïve Bayes

Naive Bayes is a simple probabilistic classifier that calculates probabilities by summing the frequencies and combinations of datasets given. The algorithm uses Bayes theorem and assumes all the independent or nonmutual attributes given by values on the class variables [15]. Naive Bayes is a classification technique with probability and statistical method brought by a British scientist Thomas Bayes, predicting the future opportunities based on the past experiences and it is known as Bayes Theorem. Combining the theorem with Naive which the condition between attributes is assuming independent. The NB classification is assumed that whether there is the presence of a certain feature or not, it has nothing to do with the characteristics of the other classes. The calculation of NB is the Xi occurrence probability in the class category of C P(C|Xi) multiplied by class category of C P(C) probability. Then the result is multiplied by the occurrence of Xi variable probability P(Xi). Mathematically, it is written in the following equation:

$$P(C \mid X_i)P = \frac{P(C \mid X_i)P(C)}{P(X_i)} \qquad (2)$$

The next process is optimal class selection by choosing the largest probability value of each class probability. Here is the formula to choose the largest value shown by (3)

$$P(C \mid X_1,\dots, X_n) = P(C)\Pi^{n}_{i=1}P(X_i|C) \qquad (3)$$

Function 3 is the Naive Bayes model which the next will be used for classification if Xi is a random variable with categorical data. If Xi is a continuous data, it is assumed as data that follow Gaus distribution with density function in (4).

$$f(x,\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (4)$$

Where μ is mean, and σ is the standard deviation.

2.2.3 Support Vector Machine

Support Vector Machine (SVM) is a leaning that uses an open space in a high dimensional feature space. Training the algorithm based on optimization theory by implementing learning bias [16]. SVM became famous because of its success in recognizing handwriting digits with 1% of errors. The basic concept of SVM is to find an optimal function that can separate two datasets for two different classes. This technique has a convincing performance in predicting a new data class.

SVM is in the same class with the Artificial Neural Network, which is including in the supervised learning, but in its implementation, SWM gives better results than ANN, especially in achieving the solutions. SVM has a good performance for solving many problems of identification [17]. Moreover, SVM can find the optimum solution in each running [18]. According to [19], the SVM method is efficient to solve classification for binary class.

The maximum margin hyperplane gives the maximum separation between the decision classes as shown in Figure 1. If the training dataset is an imbalance, then the choice of the optimal hyperplane was affected dominantly by samples vectors of majority class, a class which has much more samples data [13]. The separator function to determine the data class for x is as (5):

$$f(x) = x.w + b \qquad (5)$$

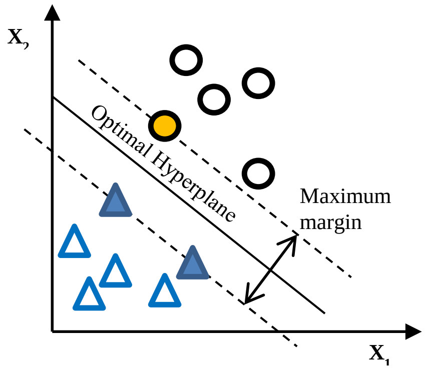

Figure 1. The maximum margin hyperplane of SVM

Where w and b are coefficients that estimated by minimizing the regularized risk function. Kernel method is the solution that uses to duplicate SVM when the data is hard or maybe impossible to be classified with limited linear fields. The use of Kernel method caused a data x in input space being mapped in the F feature space with higher dimensional by φ map as φ : x → φ(x). This mapping is doing to keeping the data characteristics or data topology. Some of the Kernel general forms that used for the SVM method are linear, polynomial, radial basis function, and sigmoid.

2.2.4 Random Forest

Random forest is one of the ensemble methods to improve the accuracy of a data classification of an unstable single divider through multiple combinations of methods similar to the voting process to get final classification prediction. The term RF was proposed by [20] from Bootstrap Aggregating process or more popularly known as Bagging. In the bagging process, bootstrap resampling is used to generate a classification tree. The classification tree is a general technique with multiple versions which then it combines to obtain the final prediction. Where the RF method, randomization process is not only done on the sample data but also on the independent variables collection, so that the classification tree raised, will have the different sizes and shapes. RF is a development of a decision tree (DT). In the DT, the classification tree is made in only one, while in RF is made more than one and it overcomes noise and missing value. The algorithm of RF is shown by:

Step 1: To get training data, generate new random sample with bootstrap resampling method N times.
Step 2: Make the decision tree or regression tree based on data by Step 1
Step 3: Repeat Step 1 and Step 2, so it will obtain several trees and become a forest
Step 4: Let each of the trees choose the X
Step 5: Count the number of the chosen X_i in each class. The class with the most number is the determinant of a classification label for $X_{i.}$
Step 6: The improper percentage classification is the class error ratio in the random forest.

According to [21], in the imbalanced prediction using random forests, there are two approaches: one is cost sensitive learning which incorporates class weights into the random forests classifier, and the other is by using over-sampling methods with the minority class and or under-sampling with the majority one to balance the original data.

2.2.5 Accuracy Measurement

Classification accuracy is used to assess the goodness of a model in representing or classifying actual events. The measure of classification accuracy used for unbalanced data is Area Under ROC Curve (AUC). AUC is complete accuracy in the context of imbalance accuracy. In performing AUC calculations, it needs to calculate sensitivity and specificity first. For easier calculations, it usually uses a confusion matrix. The formula to calculate the sensitivity, specificity and AUC score is shown by (6), (7), and (8).

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{FP}{FP + TN} \tag{7}$$

$$AUC = \frac{1 + Sensitivity - Specificity}{2} \tag{8}$$

There are other classification evaluation measures; there are Geometric Means (G-means) which was introduced by [22]. The basic idea was to maximize the accuracy of each class by keeping the balance of the both.

$$G - means = \sqrt{sensitivity * specifivity} \tag{9}$$

Study by [6] were using F-measure to evaluate the classification accuracy on the imbalance class dataset. F-measure is a combination of sensitivity and specificity which it is used to determine the best prediction result.

$$F - measure = \frac{TP}{TP + FP} \tag{10}$$

3. RESULTS AND ANALYSIS

3.1. Data Balancing

Before doing a classification analysis using NB, SVM, and RF, it is essential to know the description of the data used. In the research, the methodology has been explained that in the pre-processing stage, the data will be divided into training and testing data. In this study, the proportion of sample training and testing data is divided into four, i.e., 90:10 which means 90% of training data and 10% of testing data, 80:20, 70:30 and 50:50. It is done to get the best proportion information by applying this proportion to the original data and the data after SMOTE. Table 1 presents a summary of the percentage of data testing for the negative and positive classes in each combination.

After determining the sampling proportion, performing classification on the data with three methods of Naive Bayes, SVM, and Random Forest to evaluate the classification method. Measuring the goodness of the methods is using evaluation classification, i.e., accuracy, AUC, F-measure, and G-means. Table 2 shows the accuracy of each sample proportions of the original data and after SMOTE data the accuracy of each methods are presented in the Table 3.

Table 1. Class Distribution based on Sampling Proportion

Data	90:10		80:20		70:30		50:50	
	Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
Original	88.05%	11.95%	89.05%	10.95%	88.35%	11.65%	88.36%	11.64%
After SMOTE	78.97%	21.03%	79.17%	20.83%	78.86%	20.62%	79.45%	20.55%

Table 2. The Accuracy of Each Classifier using the Original and After SMOTE Data

Data		Original			After SMOTE	
combination	NB	SVM	RF	NB	SVM	RF
90:10	85.6%	89.2%	89.4%	83.5%	89.35	91.1%
80:20	87.5%	89.7%	89.9%	83.7%	88.2%	89.2%
70:30	87.8%	89.4%	89.7%	84%	88.4%	89.3%
50:50	87.9%	89.8%	89.5%	83.5%	88%	89.4%

Table 3. The AUC Score of Each Classifier using the Original and After SMOTE Data

Data			Original			After SMOTE
combination	NB	SVM	RF	NB	SVM	RF
90:10	67%	61.8%	61.2%	74.4%	79%	82.2%
80:20	71.3%	61.9%	62.5%	75.7%	77.8%	78.5%
70:30	68.4%	60.8%	60.9%	76.3%	78.6%	80%
50:50	70%	61.5%	59.4%	78.4%	78.6%	79.7%

3.2. Comparison of classifier

Comparison of classification accuracy for each classifier were evaluated by some measurement. Table 2 shows that the highest accuracy value on the NB method is to use 50:50 sampling proportion in the original data with accuracy 87.9%. Likewise on the SVM method with accuracy equal to 89.8%. For the RF method, the highest accuracy is obtained through the 90:10 sampling proportions in the SMOTE data. However, accuracy is considered inappropriate to be used as an evaluation of the goodness of the classification model on the dataset with an unbalanced class. It is because of the accuracy formulation based on accurate observation in the negative and positive class.

Table 3 shows that the AUC obtained by NB method with 50:50 sampling proportions of sampling data is the largest among another sampling proportions. The highest AUC score with the SVM method was 79% which is obtained by 90:10 sampling proportions with data after SMOTE, as well as RF method with AUC value equal to 82,2%. In Table 3 it can also be seen that the AUC in the SMOTE data tends to have a more significant value than the original data. Based on the method, the most considerable value is obtained by RF with a 90:10 sampling proposal of the data after SMOTE.

Based on Table 4, it can be known that the highest F-measure among three methods is in the SMOTE data with 90:10 sampling proportions. F-measure is one of the evaluation measures that appropriate for data with imbalance class, the higher the F-measure, the better the classification method, because F-measure is obtained by classification observation accuracy in the positive class only.

The last evaluation measurement is G-means, the analysis result is presented in Table 5. Table 5 shows that the largest G-means is obtained by the NB method and it is equal to 88,2% by 80:20 proportions sampling in the original data. For SVM method is, the G-means score is equal to 76,9% with 50:50 sampling proportions in the SMOTE data and the highest G-means score by RF method is obtained by 90:10 sampling proportions in the SMOTE data.

Based on the proportion samplings of training and testing data, most classification evaluation measurement in three methods obtain the highest value at 90:10 proportions. It means that the larger sample used to generate the classification model, it will describe the unbalanced data conditions. So that when using testing data for validation it obtains high AUC, F-measure, and G-means score. Besides, if the original data is compared with the data after SMOTE, the analysis results denote that SMOTE data performance is better than the original data. It matches with the theory and previous research which once states that the SMOTE sampling is used to solve the class imbalance so that the classification evaluation obtained is appropriate.

Table 4. The F-measure Score of Each Classifier using the Original and After SMOTE Data

Data	Original			After SMOTE		
combination	NB	SVM	RF	NB	SVM	RF
90:10	40.4%	60.8%	65.0%	61.4%	83.3%	87.7%
80:20	43.9%	56.5%	58.7%	60.8%	73.0%	82.5%
70:30	47.6%	61.7%	66.1%	60.8%	77.2%	80.0%
50:50	47.9%	66.3%	65.4%	58.1%	75.0%	81.1%

Table 5. The G-means Score of Each Classifier using the Original and After SMOTE Data

Data	Original			SMOTE		
combination	NB	SVM	RF	NB	SVM	RF
90:10	62.4%	50.3%	48.6%	72.6%	77.0%	80.8%
80:20	88.2%	50.6%	51.6%	74.4%	75.7%	76.4%
70:30	63.5%	47.9%	48.0%	75.1%	76.8%	78.4%
50:50	66.1%	49.3%	44.6%	78.0%	76.9%	77.9%

The best method is determined by calculating the average of all evaluation measurements of the data after SMOTE sampling proportions on each evaluation measure. The most extensive evaluation measurement will be selected as the best method. Figure 2 shows that the method with the highest average of AUC, F-measure, and G-means value among another method is RF. Therefore, the best method for this study is RF with the data after SMOTE.

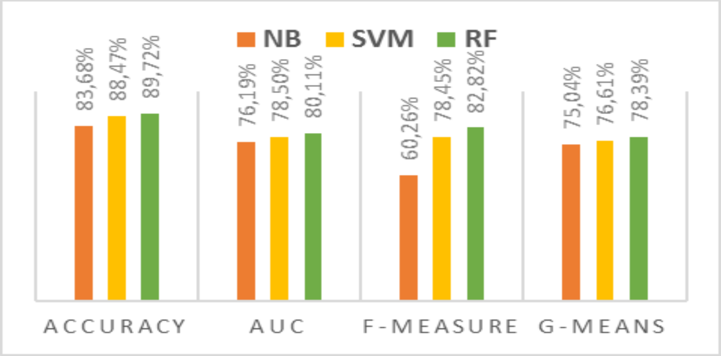

Figure 2. Comparison performance of each classifier

4. CONCLUSION

Based on the analysis, we conclude that data after resampling by SMOTE obtained a better performance than original data. This research has accomplished the objectives where three classifiers (NB, SVM, and RF) were performed for an imbalanced class dataset. The primary objective of this study is to identify the best technique for imbalanced class prediction before and after resampling by SMOTE. Hence, after applying the three methods, a comparative analysis has been performed to determine the most appropriate one. The experimental results showed that RF performs well because of its abilities to predict the higher portion of data with higher AUC, F-measure, and G-means score.

For future work, the following suggestions can be considered; Combining other resampling methods such as Tomek links and Random Under-sampling; Use more sample imbalanced dataset with a different distribution of class would be a valuable idea.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support from a Head of Institute for Research and Community Service ITS through Research Grant for Beginner Researcher scheme in 2018.

REFERENCES

[1] Phoungphol P. A Classification Framework for Imbalanced Data. Georgia State University. 2013.

[2] Rohini RR, Krishnamoorthi M. Learning from a Class Imbalanced Public Health Dataset: A Cost-based Comparison of Classifier Performance. *International Journal of Electrical and Computer Engineering (IJECE)*. 2017 ;7(4): 2215-2222.

[3] Qiang W. A Hybrid Sampling SVM Approach to Imbalanced Data Classification. Abstract and Applied Analysis. 2014; 1: 1-7.

[4] Sukarda B, Muhammad MI, Xiu Y, and Kazuyuki M. MWMOTE – Majority Weighted Minority Oversampling Technique for Imbalanced Dataset Learning. IEEE Transactions on Knowledge and Data Engineering. 2014; 26(2): 405–425.

[5] Giovanna M, Nicola T. Training and Assessing Classification Rules with Imbalanced Data. Data Mining and Knowledge Discovery. 2014; 28(1): 92–122.

[6] Galar M, Fernandez A, Barrenechea E and Herrrera F. EUSBoost : Enhancing ensembles for highly imbalance data-sets by evolutionary undersampling. *Pattern Recognition*. 2013: 3460-3471.

[7] Choi MJ. A Selective Sampling Method for Imbalanced Data Learning on Support Vector Machines. Graduate Theses. US: Iowa State University; 2010.

[8] Yap BW, Rani KA, Aryani H, Rahman A, Fong S, Khairudin Z and Abdullah NN. *An Application of Oversampling, Under-sampling, Bagging and Boosting in Handling Imbalanced Datasets*. Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Stanford. 2015; 285: 13–23.

[9] Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence and Research*. 2002; 16: 321-357.

[10] Sain H and Purnami SW. *Combine sampling support vector machine for imbalanced data classification*. Proceeding of The Third Information Systems International Conference. Surabaya. 2015; 72: 59-66.

[11] Maira A and Mohsin A. Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets. European Scientific Journal. 2017; 13(33): 340-353.

[12] Manal A, Mouaz A, Steven K, Clinton B, Jonathan E, and Sherif S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS ONE. 2017; 12(7): e0179805.

[13] Kusuma WA, Noviana N, Hasibuan LS, Nurilmala M. Improving DNA Barcode-based Fish Identification System on Imbalanced Data using SMOTE. *TELKOMNIKA (Telecommunication Computing Electronics and Control).* 2017; 15(3): 1230-1238.

[14] Lokesh SK and John SU. Comparative Study of Recommendation Algorithms and Systems using WEKA. International Journal of Computer Applications. 2015; 110(3).

[15] Patil TR and Sherekar SS. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science and Applications*. 2013; 6(2): 256-261.

[16] Vapnik VN. Support-vector networks. *Machine Learning*. 1995; 20: 273-297.

[17] Batuwita R and Palade V. *Efficient resampling methods for training support vector machines with imbalanced datasets*. Proceeding of International Joint Conference on Neural Networks. Barcelona, Spanyol. 2010: 1-8

[18] Seiffert C, Khoshgoftaar TM, Hulse JV and Napolitano A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybernet*. 2010; 40: 185-197.

[19] Miner G, Nisbet R, Elder J, Delen D and Fast A. Practical Text Mining and Statistical Analysis for Unstructured Text Data Applications. First Edition. USA: Academic Press. 2012: 1000.

[20] Breiman L. Random forests. *Machine Learning*. 2001; 45(1): 5-32.

[21] Zhou L, Wang H. Loan Default Prediction on Large Imbalanced Data Using Random Forests. *TELKOMNIKA (Telecommunication Computing Electronics and Control).* 2012; 10(6): 1519-1525.

[22] Kubat M and Matwin S. Addressing the Curse of Imbalanced Training Set: One Sided Selection. *Proceeding of the 14th International Conference on Machine Learning*. Nashville, USA. 1997: 179-186.