The Study of Safe-Level SMOTE Method in Unbalanced Data Classification

Qorry Meidianingsih, Erfiani, Bagus Sartono

Abstract— Class imbalanced data is a condition that the number of observations in one class is much greater than the other class. SMOTE method was known as a pioneer in dealing with balanced data issues. One of the methods which inspired by SMOTE is safe-level SMOTE. This study examines the safe-level SMOTE method by applying it to the various types of simulation data which built on the proportion of data imbalances, the number of nearest neighbors, and the position of minority to the majority data on the scatter plot diagram. There were three basic positions of minority to the majority data, such as separated, intersected, and overlaid. The research conclusions were obtained based on paired t-test for F-measure which generated by the prediction process using radial basis kernel SVM. The results showed that in the some kind of unpartitioned minority class instances, such as intersected and overlaid, safe-level SMOTE had a better performance than SMOTE method generally. In contrast, when the minority class instances were separated to the majority class then the both methods had a same performance. In the case of partitioned minority class instances, when the the greater proportion of minority class instances were positioned to be intersected or overlaid to the majority class instances, safe-level SMOTE method showed a better performance than SMOTE method. When the greater proportion of minority class instances were positioned separatedly to the majority class instances then the decisions to accept H0 were increasing.

Index Terms— unbalanced data, SMOTE, Safe-level SMOTE, Support Vector Machine, F-measure

1 Introduction

Data mining is a series of processes to reveal the hidden information in a large data sets. The information was obtained by extracting and recognizing a pattern of data which contained in it [7]. The application of data mining has been used in many fields, such as *Customer Relationship Management* (CRM), fraud detection of credit card, prevention of terrorism, and others. One method of the data mining is predictive method, the process of summarizing and grouping the data with the respon variable. It is the most applicable and profitable method. Classification is one technique that is used in predictive method [8].

The problems of classification had been learned by data mining and machine learning communities in various fields, such as text, multimedia, social network, and biology. Classification is a diverse topic and the algorithms used in it was very depend on data domain and problem scenario. That is why the problems that may be found are more diverse [1]. One important issue that being studied by researcher is the problem of class imbalanced data. Class imbalanced data is a condition that the number of observations in one class is much greater than the other class.

The ordinary classification method can not give the right decision when the data is unbalanced. It is caused by one class had more instances so that the classification result will be biased or generate the misclassification. The approach of dealing with unbalanced data consist of two solutions. The first solution is to make the distribution of data more balanced. It commonly known as the level of data preprocessing. The second solution is to modify the algorithm of the ordinary classification method. It can be used without changing the data distribution. At the level of data preprocessing, we can remove some instances in majority class (undersampling) or add some instances in minority class (oversampling). The method of undersampling may remove the valid instances which provide the important information. Akbani *et al* revealed that when the undersampling method performed in majority class then the instances is no longer random [2]. Batista *et al* compared some methods of undersampling and oversampling to overcome the unbalaced data problem [3]. In general, based on ROC curve, oversampling method showed the better results than the undersampling method.

Synthetic Minority Over-sampling Technique (SMOTE) is a popular oversampling technique and proposed by Chawla at 2002 [6]. The basic idea of SMOTE is to generate the synthetic data along the line between minority instances and its nearest neighbours. The weakness of SMOTE is ignoring tha area around the synthetic instances. It causes the synthetic instances may be generated around the majority instances so that misclassification can be occurred. One method that can handle that problem is Safe-level SMOTE. This method proposed by Bunkhumpornpat *et al* at 2009 [4]. The principle of circumspection can be seen in this method. The basic idea of Safe-level SMOTE is to generate the synthetic instances in a safe area. The criteria for the safe area is based on a coefficient called *safe-level ratio*. Given these criteria, the synthetic in-

- *Qorry Meidianingsih is currently pursuing masters degree program in statistics in Bogor Agriculture University, Indonesia, PH +6281297142104. E-mail: qorry88@gmail.com*
- *Erfiani is Lecturer, Departement of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: erfiani_ipb@yahoo.com*
- *Bagus Sartono is Lecturer, Departement of Statistics, Bogor Agriculture University, Bogor, Indonesia. E-mail: bagusco@gmail.com*

stances were expected to be located among the minority area. Therefore the oversampling method works more effective and the result is more appropriate.

This study examines the Safe-level SMOTE method by applying it to the various types of simulation data which built on the proportion of data imbalances, the number of nearest neighbors, and the position of minority to the majority data on the scatter plot diagram. Based on these simulation design, the new information on how well the Safe-level SMOTE method works is expexted to be obtained.

2 Research Method

2.1 Data

Data used in this study were obtained by simulation processes with two variables, X1 and X2. The respon variables (class) consist two categories namely 1 and 2. Class 1 is a majority class while class 2 is a minority class. The position of minority instances in scatter plot diagram was consisted to be unpartitioned (P_0) and partitioned into two parts while the majority instances were unpartitioned .

TABLE 1
COMPARISONS OF PARTITIONED MINORITY OBSERVATION NUMBER

Symbol	Partition 1	Partition 2
P_0	m	-
P_1	10	m-10
P_2	50%m	50%m
P_3	10%m	90%m
P_4	25%m	75%m
P_5	40%m	60%m

The "m" value describe the number of minority class observations.

Any minority data set, either unpartitioned or partitioned, will be positioned separately, intersectedly, and overlaidly to the majority data set. In separated position there is a long enough distance between the majority and minority data. It was indicated by the value of $\boldsymbol{\mu}$ parameter which was much different between the two classes. Intersected position indicate that theres is some observations from the both class which have the same coordinate in scatter plot diagram. Therefore the value of $\boldsymbol{\mu}$ parameter between the class is not much different. When the position of the observations from the both class is overlaid then the value of $\boldsymbol{\mu}$ parameter is same.

The three types of the basic position then combinated each other so that there are 40 types of simulation data. The types of simulation data is shown in Table 2. The number of population observations was determined by 10000 and it was generated by normal multivariate distribution. The population was divided based on the proportion of data imbalances, there was Population 1 (95%:5%), Population 2 (90%:10%), and Population 3 (85%:15%).

TABLE 2
THE TYPES OF SIMULATION DATA

Minority Data Position	The number of minority class observation					
	P_0	P_1	P_2	P_3	P_4	P_5
Separated (S)	√	-	-	-	-	-
Intersected (I)	√	-	-	-	-	-
Overlaid (O)	√	-	-	-	-	-
Intersected - Intersected (I-I)	-	√	√	√	√	√
Intersected - Overlaid (I-O)	-	√	√	√	√	√
Separated - Intersected (S-I)	-	√	√	√	√	√
Separated - Overlaid (S-O)	-	√	√	√	√	√
Separated - Separated (S-S)	-	√	√	√	√	√
Intersected - Separated (I-S)	-	√	-	√	√	√
Overlaid - Intersected (O-I)	-	√	-	√	√	√
Overlaid - Separated (O-S)	-	√	-	√	√	√

Symbol "-" means simulation was not conducted.

2.2 Methods of Data Analysis

The step of analysis in this study are follows.

Preparation of Simulation Data

1. Generate the simulation data for population 1, 2, and 3 based on the distribution criteria shown in Table 2.
2. Create the scatter plot diagram for each population. It can be used to make sure that the generated data as expected before.
3. Select sample of 10% of population data by stratified random sampling technique that the allocation sample in each strata was proportional. In this study, majority and minority class were determined to be strata because of the observation characteristic differences.
4. Create the scatter plot diagram for each sample. It can be used to make sure that the sample distribution was approximately identical with population distribution.
5. Divide the sample data to be training set (80%) and test set (20%).

Application of SMOTE Method

1. Calculate the distance between the minority class instances in training set and its nearest neighbors using Euclidean distance.

$$\Delta(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

   The neighbors were coming only from minority class instances.
2. Take the k nearest neighbours. In this study, the value of k are 5, 15, and 30.
3. Randomly select one of k nearest neighbors (n).
4. Calculate the difference between the minority class instances in training set and its nearest neighbor chosen in step 3.

5. Multiply the difference obtained in step 4 by a random number between 0 and 1.
6. Add the result of step 5 to the minority class instances in training set. That is the new instance.
7. Repeat the steps above until the the number of minority class observations were approximately similar with the number of majority class observations.

Application of Safe-level SMOTE Method

1. Determine the criteria area to generate the synthetic data. The step are follows.
   1) Calculate the Euclidean distance between the minority class instances in training set (p) and its nearest neighbors.

$$\Delta(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

   The neighbors were coming from minority and majority class instances.
   2) Take the k nearest neighbours. In this study, the value of k are 5, 15, and 30.
   3) Randomly select one of k nearest neighbors (n) that come from the minority class.
   4) Recalculate the distance between n and its neighbors with the same k using Euclidean distance.
   5) Randomly select one of k nearest neighbors (n) that come from the minority class.
   6) Calculate the safe-level to p dan n.
      Safe-level (p) : the number of minority class instances in k nearest neighbors for p
      Safe-level (n) : the number of minority class instances in k nearest neighbors for n
   7) Calculate the safe-level ratio for p and n.
      Safe-level ratio : Safe-level (p) / Safe-level (n)
2. Generate the synthetic data based on safe-level ratio [4].
   1) Calculate the difference between p and n.
   2) Take the range of random number based on the safe-level ratio obtained.
      - SLR = ∞ and Safe-level (p) = Safe-level (n) = 0
        In this case p and n are noise so that there is no synthetic data be generated.
      - SLR = ∞ and Safe-level (p) $\neq$ 0
        In this case n is noise. The synthetic data will be generated far from n by duplicating p.
      - SLR = 1
        The synthetic data will be generated along the line between p and n because p is as safe as n.
      - SLR > 1
        In this case the safe-level of p is greater than safe-level of n. The synthetic data will be generated closer to p at distance [0,1/SLR].
      - SLR < 1
        In this case the safe-level of n is greater than safe-level of p. The synthetic data will be generated closer to n at distance [1-SLR,1]
   3) Multiply the difference obtained in step 1 by a random number obtained in step 2 (only for case 3 to 5).
   4) Add the result of step 3 to p. That is the new instance.
   5) Repeat the steps above until the the number of minority class observations were approximately similar with the number of majority class observations.

Evaluation of *Oversampling* Method

1. Applicate the Support Vector Machine method with the kernel radial basis used (gamma=0.5) to the new data.
2. Predict the test set using classification model obtained in step 1.
3. Evaluate the performance of classifier by calculating the F-measure.

$$\text{F-measure} = \frac{(1+\beta^2) \times \text{recall} \times \text{precision}}{\beta^2 \times \text{recall} + \text{precision}}$$

   where

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

   True Positives (TP) is the number of the positive class instances which correctly classified to the positive class. False Positives (FP) is the number of negative class instances which classified to the negative class (incorrectly classified). False Negatives (FN) is the number of positive class instances which classified to the negative class (incorrectly classified). β indicate the relative importance between precision and recall and it was set to be 1 [5].
4. Compare the F-measure obtained by Safe-level SMOTE method and SMOTE method using paired t test.

The step of analysis which include obtained the sample data until calculated the F-measure were conducted repeatedly until 100 times.

3 Result and Discussion

3.1 Illustration of Simulation Data

Fig 1 illustrates some scatter plot diagrams of simulation data. The red triangle symbol describes the instances of minority class while the other one describes the instances of majority class. Scatter plot (a), (c), and (e) showed the simulation data of the unpartitioned minority class observations such as separated, intersected, and overlaid with the number of minority class observation m.

On the other side, scatter plot (b), (d), and (f) showed that the minority class instances were partitioned into two parts. Scatter plot (b) showed the separated-intersected (S-I) position

of minority class instances to the majority class instances. The type of observation number used was P_1.

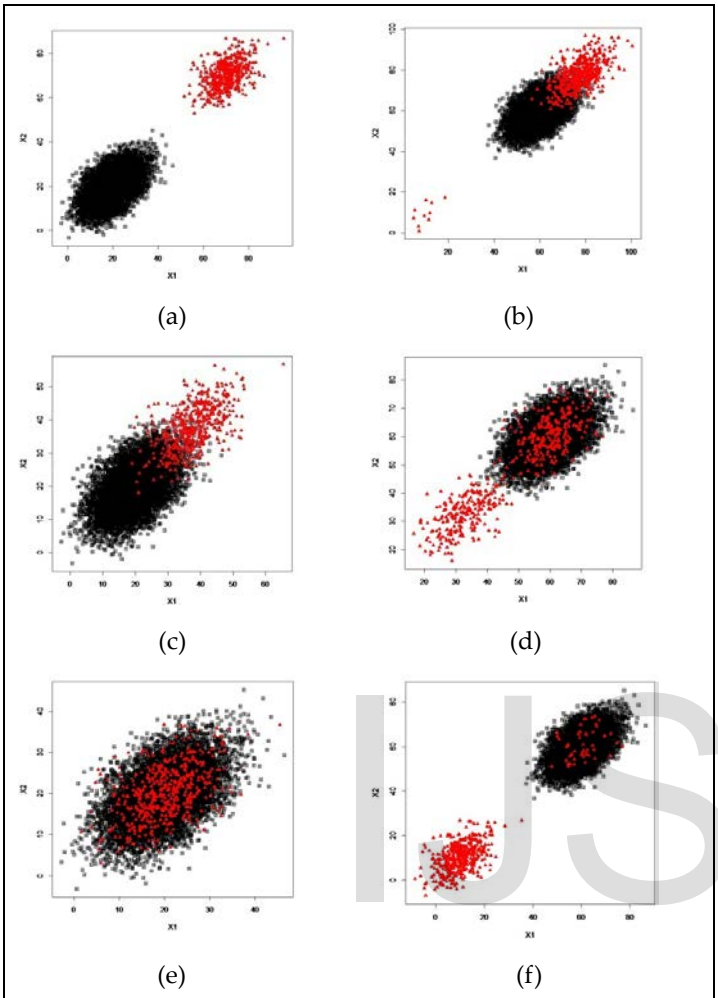

Fig 1. Ilustrations of simulation data distribution of (a) separated, (b) intersected, (c) overlaid, (d) separated- intersected, (e) intersected-overlaid, (f) overlaid-separated.

In separated position, the number of minority class instances were 10 while the remain instances (m-10) was positioned intersectedly with the majority class instances. Scatter plot (d) showed the intersected-overlaid (I-O) position of minority class instances to the majority class instances with the type of observation number used was P_2. Therefore the number of observation of both class was equal (50%m). The last scatter plot (f) showed the overlaid-separated (O-S) position of minority classs instances to the majority class instances. Type of observation number used was P_3. In overlaid position there were 10%m instances of minority class and the the remain instances (90%m) were positioned separatedly from the majority class instances. The m is the number of minority class observation. It has a various amount based on the type of population. In this illustrations, the type of population that used is population 1.

3.2 Evaluation of Oversampling Method

The first step will be conducted was testing the performance of both method based on F-measure using paired t test. When H_0 is rejected then the both method had the different performance. Finally the best performance method was determined by the greater F-measure. Table 3 summarizes the results of hypothesis testing (α=5%).

TABLE 3
THE RESULTS OF F-MEASURE TEST USING PAIRED T TEST

Position	Population	Type of the partitioned minority observation number																	
		P_0			P_1			P_2			P_3			P_4			P_5		
		k																	
		5	15	30	5	15	30	5	15	30	5	15	30	5	15	30	5	15	30
I	Pop 1	●	●	□															
	Pop 2	●	●	●															
	Pop 3	●	●	●															
I-I	Pop 1				●	●	□	●	●	●	●	●	●	●	●	●	●	●	●
	Pop 2				●	●	□	●	●	●	●	●	●	●	●	□	●	●	●
	Pop 3				●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
I-O	Pop 1				●	●	●	●	□	●	●	●	●	●	●	●	●	●	●
	Pop 2				●	●	●	●	□	▲	●	●	●	●	●	●	●	▲	●
	Pop 3				●	●	●	●	□	□	●	●	●	●	●	●	●	●	●
I-S	Pop 1				●	●	●				●	□	▲	●	●	▲	●	●	□
	Pop 2				□	□	□				●	□	▲	●	●	▲	●	●	●
	Pop 3				□	□	□				●	●	□	●	●	●	●	●	●
O	Pop 1	●	●	●															
	Pop 2	●	●	●															
	Pop 3	●	●	●															
O-I	Pop 1				●	●	●				●	●	●	●	●	●	●	●	●
	Pop 2				●	●	●				●	●	●	●	●	●	●	●	●
	Pop 3				●	●	●				●	●	●	●	●	●	●	●	●
O-S	Pop 1				●	●	●				●	●	●	□	●	●	●	□	□
	Pop 2				□	□	□				□	●	●	□	□	●	□	□	●
	Pop 3				□	□	□				□	□	●	□	□	□	□	□	□
S	Pop 1	□	□	□															
	Pop 2	□	□	□															
	Pop 3	□	□	□															
S-I	Pop 1				●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
	Pop 2				●	●	□	●	●	●	●	●	●	●	●	●	●	●	□
	Pop 3				●	●	●	●	●	●	●	●	●	●	□	●	●	●	□
S-O	Pop 1				●	●	●	●	□	□	●	●	●	●	●	●	●	□	□
	Pop 2				●	●	●	□	▲	▲	●	●	●	●	□	●	●	▲	□
	Pop 3				●	●	●	□	□	□	●	●	●	●	□	●	□	▲	□
S-S	Pop 1				□	●	□	□	□	●	●	●	●	□	●	●	□	□	●
	Pop 2				●	□	□	□	□	□	□	●	●	□	□	●	□	□	□
	Pop 3				□	□	□	□	□	□	□	□	●	□	□	□	□	□	□

Symbols in the table above characterized the results; □ = both method had same performance, ● = Safe-level SMOTE is better than SMOTE, ▲=SMOTE is better than Safe-level SMOTE, the grey patterns mean simulation was not conducted.

In the case of unpartitioned minority class observations, such as intersected, Safe-level SMOTE had a better performance than SMOTE method generally. It was similar with the results of overlaid position. The separated position had a different results. In this position, there was no different perfomace by the both method. It was clear that when the minority class instances was separated to the majority class instances

then the synthetic data was generated in the safe area.

When the all partitioned minority class instances was intersected to the majority class instances (I-I) there was 3 types of simulation data showed the results that the both method had a same performance. They are population 1 (P_1) and population 2 (P_2 and P_5). The similarity of these simulation data was that the value of *k* used, i.e. 30. In the case that the minority class instances were partitioned to be intersected and overlaid (I-O) to the majority class instances, the obtained results were not much different as the (I-I) type. It showed that the Safe-level SMOTE method had a better performance than SMOTE in most of all data types, but the P_2 criteria gave a contrast result at *k*=15. When the the minority class instances were pastitioned to be intersected and separated to the majority class instances (I-S), the results of the method had a same performance were more obtained than the I-I and I-O obtained. Population 2 and 3 showed it consistently at P_1 criteria for each *k*.

The minority class instances were partitioned to be overlaid and intersected (O-I) showed the very consistent results for all the criteria. All of these data types showed that the Safe-level SMOTE method had a better performance than SMOTE method. The overlaid and intersected positions were quite risky for the occurence of classification error so that the synthetic data must be generated carefully. In the overlaid-separated (O-S) position, the smaller amount of minority class instances were located overlaidly to the majority class instances while the most of the others were separated. It caused most of the synthetic data were generated in the separated position so the probability of missclassification was very small. Therefore the Safe-level SMOTE and SMOTE method had a not different performance in dealing with unbalanced data problem for this kind of simulation data.

In the case that the minority class instances were partitioned to be separated and intersected to the majority class instances (S-I), Safe-level SMOTE method had a better performance than SMOTE method. It was contrast with the results of the minority class instances were partitioned to be separated-intersected (S-I). In that case, the decision that the both methods were equally good was more obtained, moreover the SMOTE method had a better performance in some kind of simulation data. The results obtained in the case that the all partitioned minority class instances was separated to the majority class instances (S-S) seem to be predictable. The Safe-level SMOTE method had a performance that was not different with the SMOTE method for most kind of this simulation data.

4 Conclusion

Based on the results of testing the F-measure, in the some kind of unpartitioned minority class instances, such as intersected and overlaid, Safe-level SMOTE had a better performance than SMOTE method generally. In contrast, when the minority class instances were separated to the majority class then the both method gave the same performance.

In the case of partitioned minority class instances, when the the greater proportion number of minority class instances were positioned to be intersected or overlaid to the majority class instances, Safe-level SMOTE method showed a better performance than SMOTE method. The different results were obtained when the greater proportion number of minority class instances were positioned separatedly to the majority class instances. In the S-S, O-S, and I-S criteria, the decision that both method had a same performance was more appear. It because the separated position allows the synthetic data were generated to be far enough from the majority class instances.

References

[1] Aggarwal CC. 2015. *Data Classification : Algorithms and Applications*. CRC Press, Taylor & Francis Group, LLC.

[2] Akbani R, Kwek S, Japkowicz N. 2004. Applying Support Vector Machines to Imbalanced Datasets. *Proceedings of the 15th European Conference on Machine Learning*. pp. 39-50.

[3] Batista GEAPA, Prati RC, Monard MC. 2004. *A study of the behavior of several methods for balancing machine learning training data*. SIGKDD Explorations, 6(1).

[4] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. 2009. *Safe-Level-SMOTE:* Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem. In : *Theeramunkong T, Kijsirikul B, Cercone N, & Ho T-B. (eds). PAKDD. LNCS, Vol. 5476, pp. 475-482*. Springer, Heidelberg.

[5] Chawla VN. 2005. Data mining for imbalanced datasets : an overview. In : Oded Maimon & Lior Rokach, editors. *The Data Mining and Knowledge Discovery Handbook*. US : Springer. 853-867.

[6] Chawla VN, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 16 : 321-357.

[7] Han J, Kamber M. 2006. *Data Mining : Concepts and Techniques, 2nd edition*. San Fransisco : Morgan Kaufmann.

[8] Tuffery S. 2011. *Data Mining and Statistics for Decision Making*. UK : John Wiley & Sons, Ltd.