Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review

B Santoso[1], H Wijayanto[1], K A Notodiputro[1] and B Sartono[1]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Indonesia.

Email : budsant.bs@gmail.com, hari_ipb@yahoo.com, khairilnotodiputro@gmail.com, bagusco@gmail.com

Abstract. Class imbalanced commonly found in any real cases. Class imbalanced occur if one of the classes has smaller amount, called minority class, than other class (majority class). The problem of imbalanced data is usually associated with misclassification problem where the minority class tends to be misclassified as compared to the majority class. There are two approaches should be performed to solve imbalanced data problems, those are solution at data level and solution at algorithm level. Over sampling approach is used more frequently than the other data level solution methods. This study gives review of synthethic over sampling methods for handling imbalance data problem. The implementation of different methods will produce different characteristics of the generated synthetic data and the implementation of appropriate methods must be adapted to the problems faced such as the level and pattern of imbalanced data of data available. Results of the review show that there is no absolute methods that are more efficient in dealing with the class imbalance. However, the class imbalance problem depends on complexity of the data, level of class imbalance, size of data and classifier involved. Determination of over sampling strategy will affect the outcome of the over sampling. So it is still open better development oversampling methods for handling the class imbalance. The selection classifier and evaluation measures are important to get the best results. Statistical test approach is needed to assess the theoritical propertis of synthetic data and evaluate missclassification in addition to the evaluation methods that have been used.

1. Introduction

Class imbalanced commonly found in any real cases. Class imbalanced occur if one of the classes has smaller amount, called minority class, than other class (majority class). The problem of imbalanced data is usually associated with misclassification problem where the minority class tends to be misclassified as compared to the majority class. Problem rises when data of minority class contain important information and become the focus of attention of research so that errors in classification will lead to errors in decision-making especially in prediction accuracy of minority class.

There are two approaches should be performed to solve imbalanced data problems, those are solution at data level and solution at algorithm level [1]. Solution at data level is applied by balancing the distribution of the majority and minority class through methods of under sampling, over sampling or combination of both methods. Solution at algorithm level is applied by modification in classifier methods or optimize the performance of learning algorithm. The advantage of the data level is can use independent of the classifier selected.

In under sampling process, data on majority class are reduced or eliminated some data to balance the class distribution, while over sampling process is done by adding the data on minority class. The process of balancing the distribution of data can be achieved by performing under sampling and over sampling simultaneously, termed combination or hybrid. Over sampling approach is used more frequently than under sampling since under sampling method will eliminate data in the majority class thus causing lost of important information of the data. According to Batista *et al.* [2] in general, over sampling method gives better results than under sampling method. The famous over sampling method is the Synthetic Minority Over Sampling Technique (SMOTE). However, the SMOTE method still has some weaknesses hence it motivates the research development to overcome the problem of SMOTE.

This paper will review several over sampling methods which used to solve the imbalanced data problems, those were SMOTE and development methods of SMOTE such as SMOTE-TL-ENN, SMOTE RSB, Borderline-SMOTE and Safe Level SMOTE.

2. Evaluation measures

Most of evaluation measures for two class problem are built by confusion matrix as illustrated in Tabel 1. The class label of the minority class is positive and the class label of the majority class is negative. The first row of the table is the actual class label and the first column present their predicted class label. TP and TN denote the number of positive and negative examples that are classified correctly, while FN and FP denote the number of misclassified positive and negative examples respectively.

Table 1. Confusion matrix.

		Predicted class	
		Yes	No
Actual Class	Yes	TP: True positive	FN: False negative
	No	FP: False positive	TN: True negative

To evaluate class imbalanced, classification accuracy is not sufficient as a standard performance measure. Classification accuracy can not be used because it does not specifically describe the classification of a particular class. Receiver Operator Characteristics (ROC) analysis, Area Under the ROC Curve (AUC) and metrics such as precision, recall and F-value more better to understand the performance on the minority class.

Precision, recall and F-value can be expressed as:

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

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - value = \frac{(1 + \beta^2)\ Recall\ x\ Precision}{\beta^2\ Recall + Precision}$$

Where β corresponds to relative importance of precision vs. recall and it is usually set to 1.

The goal of evaluation in class imbalance is to improve the recall, without sacrificing the precision. However, the recall and precision goals are often conflicting and attacking them simultaneously may not work well, especially when one class is rare. The F-value incorporates both precision and recall, and the "goodness" of a learning algorithm for the minority class can be measured by the F-value. While ROC curves represent the trade-off between values of TP and FP, the F-value basically incorporates the relative effects/costs of recall and precision into a single number [3]. ROC curve value under ideal conditions (0.100) indicates that all the positive samples were classified correctly and no negative sample were classified as positive sample.

3. Current approaches for handling class imbalance

3.1. Solution at data level

Solution at data level is applied by balancing the distribution of the majority and minority class through methods of under sampling, over sampling or combination of both methods.

- Under sampling
  In under sampling, data on majority class are reduced or eliminated some data to balance the class distribution. The simple approach of under sampling method namely RUS (Random Under Sampling) where data on the majority of reduced class randomly. Furthermore, various developing methods under sampling conducted by Tomek [4] is by way of modification of Condensed Nearest Neighbor (CNN) and Tomek-Link. Furthermore Kubat and Matwin [5] using the method of One Side Selection (OSS). Laurikkala [6] proposed a new method, namely Neighborhood Cleaning Rule (NCR) which proved to be better than simple random and OSS. Cluster-based approach proposed by Yoon [7] and followed by the Yen and Lee [8].
- Over sampling
  Procedure in over sampling by adding the data on minority class to balance class distribution. Over sampling approach is used more frequently than under sampling since under sampling method will eliminate data in the majority class thus causing lost of important information of the data. The simple over sampling method is the Random Over Sampling (ROS), that is carried out by balancing the distribution of data through the application of minority data duplication at random. However, the method has the disadvantage that ROS cause problems overfitting [9]. To solve this problem, many research on methods over sampling conducted among others Chawla *et al.* [3], Batista *et al.* [2], He *et al.* [10], Han *et al.* [11], Bunkhumpornpat *et al.* [12], Tang and Chen [13], Stefanowski and Wilk [14], Cohen *et al.* [15], Napierala *el at* [16] and Ramentol *et al.* [1].
- Hybrid
  Hybrid methods combine the under sampling and over sampling methods by eliminating the data from majority class and adding the data from minority class to balance class distribution. Research on hybrid method conducted by Barandela *et al.* [17].

3.2. Solution at algorithm level

Solution at algorithm level is applied by modification in classifier methods or optimize the performance of learning algorithm on unseen data.

- Cost sensitive learning
  Another way to improve the performance of classifier is by applying cost for minority class distribution. In solution at data level, classifier try to minimize the number or errors with new data when the cost of different errors are equal. In many real-world applications the cost of different errors are often unequal. For example in medical diagnosis, the cost of erroneously diagnosing a patient to be healthy may be much bigger than that of mistakenly diagnosing a healthy person as being sick. Research on cost sensitive learning carried out by Veropoulos [18], Ting [19] and Zhou and Liu [20].
- Ensemble
  Ensemble is a combination of multiple classifiers to improve and increase the prediction accuracy. In past few yers, ensembles have emerged as a promising technique with the ability

to improve the performance of weak classification algorithms [21]. In the field of class imbalance, ensembles have mainly been used to combine the result of several classifiers. Research using ensemble approach undertaken by Sun *et al.* [22], Freund and Schapire [21], Schapire and Singer [23], Guo and Viktor [24], Liu *et al.* [25] and Chawla *et al.* [3].

4. Selected synthetic over sampling approaches

Over sampling approach is a widely used method to solve the imbalanced data. According to Batista *et al.* [2] in general, over sampling method gives better results than under sampling method. When data is highly imbalance, significant differences between majority dan minority class can be handled by over sampling methods.

The simple over sampling method is the random over sampling, that is carried out by balancing the distribution of data through the application of minority data duplication at random. Chawla *et al.* [26] proposed a new technique, named SMOTE, to generate synthetic data based on the distance between the minority data and the closest minority data therefore the new synthetic data will be formed between the two minority data.

Although SMOTE is quite effective in improving the classification accuracy of the minority data, but there are still problems, among others, that is the occurrence of overgeneralization. Data synthetic result of SMOTE is still possible to spread on both minority and majority data, hence it will reduce the performance of classification.

Formula to generate synthetic data by SMOTE can be expressed as

$$D_{new} = D_i + \left(\widehat{D_\iota} - D_i\right) \times \delta$$

Where D_{new} = synthetic data, D_i = examples from minority, $\widehat{D_\iota}$ = one of k-nearest neighbor from D_i , δ = random number between 0 and 1

To solve the SMOTE problems, some researches are conducted to modify SMOTE in order to create more effective technique in improving the classification performance. The development of SMOTE is mostly done in two ways, the first is to improve the synthetic data results of SMOTE [2,1]. The second way is to first determine which location will be choose for data generating as it is expected that the result of data generating will exactly found in the data minority area [11,12].

Batista *et al* [2] succeeded in correcting the SMOTE data by deleting the synthetic data located in the area of minority data through the Tomek Links (SMOTE-Tomeks links). In addition to the method of Tomek Links [4], Batista *et al.* [2] also improved the synthetic data on both data (majority and minority) by using the Edited Nearest Neighbor (SMOTE-ENN). Meanwhile Ramentol *et al.* [1] used the Rough Set theory to improve synthetic data generated by SMOTE, that is the SMOTE-RSB.

Han *et al.* [11] divided three location based on the amount of majority data in the nearest neighbors of minority data. If the majority data are all around the nearest neighbors, then the area is called noisy. If the majority data around the nearest neighbors are found higher than or equal to the minority, the area is called borderline. Meanwhile if most of/all data around the nearest neighbors are the minority data, then the area is called safe. Han *et al.* [11] focused on the borderline area located on the boundary between the minority and the majority of data. Data generating through SMOTE is performed in the borderline areas, called the Borderline SMOTE.

Moreover, Bunkhumpornpat *et al.* [12] paid attention to the safe area to perform over sampling based on the ratio between the number of minority data and the nearest neighbors. This method wants to make sure that the synthetic data that will be generated through the SMOTE are in areas that are completely safe. Bunkhumpornpat *et al.* [12] defines five criteria based on the ratio of safe level. Therefore the method is called the Safe Level SMOTE.

5. Discussion

We can compare several over sampling methods by located, level of class imbalance, amount of over sampling, number of k-nearest neighbour and choosen classifiers. Based location of examples, we can divided into 3 categories: safe, borderline and noisy examples [5,11,16,1].

- Borderline examples are located in the area surrounding class boundaries, where the minority and majority classes overlap.
- Safe examples are placed in relatively homogeneous areas with minority class.
- Noisy examples are individuals from one class occurring in areas of the other class.

SMOTE, SMOTE-TL-ENN and SMOTE RSB methods using all of minority examples in whole location to create to generate synthetic data. Batista *et al.* [2] use application of Tomek Links and ENN as cleaning method over the dataset obtained by the application of SMOTE. Ramentol [1] only selects the minority synthetic data that belong to the lower approximation using Rough Set Theory until the dataset is balanced.

Borderline-SMOTE and Safe Level SMOTE only use specific location to generate synthetic data. After determining an example of a minority in a specific location and then the process of SMOTE done to generate synthetic data.

Table 2. Comparative review.

	Location	Cleaning data	Number of K-NN	Amount OS	Random number
SMOTE	Borderline, noisy and safe	No	5	various	0 to 1
SMOTE-TL-ENN	Borderline, noisy and safe	Yes	3	various	0 to 1
SMOTE RSB	Borderline, noisy and safe	Yes	5	balance	0 to 1
Borderline-SMOTE	Borderline	No	5	various	0 to 0.5
Safe Level-SMOTE	Safe	No	5	various	various

Table 2 indicate that various methods use different strategy to handle class imbalance. Several methods using number of k-nearest neighbour set to 5 except SMOTE-ENN. Nearest neighbor indicating the similarity of the data. The problem is that it could be the nearest neighbor has a considerable distance. It can lead to overgeneralization. Determination of nearest neighbors can be evaluated further by using alternatives such as radius size or percentage.

In over sampling methods, we can add data on the minority so close to the amount of data on the class of the majority or equal as the majority class. In SMOTE, Borderline-SMOTE and Safe Level methods, the minority class was oversampled at 100%, 200%, 300%, 400%, 500%. The amount of data in oversampling depend on level of imbalance. Variations in the amount of oversampling cause changes in accuracy and other measures. In SMOTE RSB method, minority data was added so that the amount of data equal to a majority data.

Based on the formula in SMOTE, the distance between the data minority and its nearest neighbors multiply by random number between 0 and 1. If the random number is close to 0 then the synthetic data will be similar to the origin minority data. Conversely, if the random number is close to 1 then the synthetic data will be similar to nearest neighbors. Problems that can occur is if the random number is about 0.5 it is possible that the synthetic data similar to the data of the majority. This is the cause of overgeneralization.

In the Borderline-SMOTE, the random number is determined between 0 to 0.5 in hopes of synthesis data will be adjacent to the minority data. Different strategies used in other oversampling methods. Safe level SMOTE use RSL (Ratio of Safe Level) to determine random number. SMOTE-TL-ENN and SMOTE-RSB use cleaning data to improve synthetic data generated by SMOTE.

Table 3. Winner of AUC Result by percentage of minority

	Winner of AUC Result*	
	≤ 5 %	>5 %
SMOTE	-	2
SMOTE-TL-ENN	3	1
SMOTE RSB	10	5
Borderline-SMOTE	2	3
Safe Level-SMOTE	5	2

*) Ramentol *et at.* [1] using 44 datasets

Table 3 shows the effectiveness of information over sampling method using several datasets based Area Under the ROC Curve (AUC). The table is divided into two levels of class imbalance, namely class imbalance below 5 percent and above 5 percent.

Based on Table 3, we can't find best method in handling the class imbalance. Ramentol *et al.* [1] using 44 datasets with several over sampling methods. As a result, there is no better method on all datasets. SMOTE only effective on 2 dataset, SMOTE-TL-ENN effective in four datasets, Borderline-SMOTE effective at 5 dataset. An effective methods in many datasets are Safe Levels-SMOTE (7 dataset) and SMOTE RSB with 15 datasets. SMOTE only effective in percentage of minority up to 5 percent, while others methods effective in any imbalance level. The same result is done by Batista *et al.* [2] which by using 16 datasets, there are not the most efficient method. SMOTE efficient at 2 dataset, SMOTE-TL efficient in four data sets and SMOTE-ENN efficient at 9 datasets.

The results of the research by Han *et al.* [11] indicate that Borderline-SMOTE more efficient than SMOTE and ROS using the F-value and Recall. However there is one dataset that showed the opposite results. The experiments by Bunkhumpornpat *et al.* [12] show that the performance of Safe-Level-SMOTE evaluated by Precision and F-value are better than that of SMOTE and Borderline-SMOTE when decision trees C4.5 are applied as classifiers.

The use of classifier method affects effectiveness. In SMOTE method, the effectiveness of different classifier methods based on the dataset used and evaluation measures. The use of C.45 method, Naive Bayes and Ripper show different efficiencies according to the different datasets and different measures matrices. The same situation also happened on research conducted by Han *et al.* [11] and Batista *et al.* [2].

Further research is needed to evaluate the classification method is more appropriate in dealing with class imbalance. Exploration statistically in the data results over sampling is required to determine the characteristics of the synthetic data is generated. Based on synthetic data characteristics can be determined more precise method used to overcome the problem of class imbalance.

6. Conclusion

Results of the review show that there is no absolute methods that are more efficient in dealing with the class imbalance. However, the class imbalance problem depends on complexity of the data (located of minority data), level of class imbalance, size of data and classifier involved. Determination of the distance and the number k-nearest neighbors will affect the outcome of the over sampling. So it is still open better development oversampling methods for handling the class imbalance. The selection classifier is important to get the best results.

In general, the evaluation measure is still much to do with the measures on confusion matrices such as accuracy, precision, recall, F-value and ROC. Statistical test approach is needed to assess the theoritical propertis of synthetic data and evaluate missclassification.

References

[1] Ramentol E, Caballero Y, Bello B and Herrera F 2011 SMOTE-RSB: A Hybrid Preprocessing Approach based on Oversampling and Undersampling for High Imbalanced Data-Sets using SMOTE and Rough Sets Theory, *Knowledge and Information Systems* (London: Springer) 245

[2] Batista GEAPA, Prati RC and Monard MC 2004 *SIGKDD Exploration* **6(1)** 20

[3] Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP 2002 *Journal of Artificial Intelligence Research* **16** 321

[4] Tomek I 1976 *IEEE Transactions on Systems, Man and Cybernetics* **6** 769

[5] Kubat M and Matwin S 1997 Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, *14th International Conference on Machine Learning (ICML97)* (USA: Tennessee) 179

[6] Laurikkala J 2001 Improving Identification of Difficult Small Classes by Balancing Class Distribution, 8th Conference on AI in Medicine in Europe AIME01 (Portugal: Cascais) 63

[7] Yoon K, Kwek S 2005 An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics, *5th International Conference on Hybrid Intelligent Systems HIS05* (Brazil: Rio de Janeiro) 303

[8] Yen S, Lee Y 2006 Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset, *International Conference on Intelligent Computing ICIC06* (China: Kunming) 731

[9] Tetko I, Livingstone D and Luik A 1995 *Chemical Information & Computer Sciences* 826

[10] He H, Bai Y, Garcia EA and Li S 2008 *IJCNN08* 1322

[11] Han H, Wang WY and Mao BH 2005 *ICIC05* 878

[12] Bunkhumpornpat C, Sinapiromsaran K and Lursinsap C 2009 *PAKDD09* **5476** 475

[13] Tang S and Chen S 2008 The Generation Mechanism of Synthetic Minority Class Examples, *5th Int. Conference on Information Technology and Applications in Biomedicine (ITAB*) (China: Shenzhen) 444-447

[14] Stefanowski J and Wilk S 2008 Selective pre-processing of imbalanced data for improving classification performance, 10th International Conference in Data Warehousing and Knowledge Discovery (Springer)

[15] Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A 2006 *Artificial Intelligence in Medicine* **37** 7

[16] Napierala K, Stefanowski J and Wilk S 2010 Learning from Imbalanced Data in Presence of Noisy and Borderline Examples, *7th International Conference on Rough Sets and Current Trends in Computing (RSCTC2010)* (Poland: Warsaw) 158

[17] Barandela R, Valdovinos RM, Sánchez JS and Ferri FJ 2004 *LNCS* **3138** 806

[18] Veropoulos K, Cristianini N and Campbell C 1999 Controlling the sensitivity of support vector machines, *16th International Joint Conferences on Artificial Intelligence IJCAI99* (Sweden: Stockholm) 281

[19] Ting K M 2002 *IEEE Transactions on Knowledge and Data Engineering* **14(3)** 659

[20] Zhou Z, Liu X Y 2006 *IEEE Transactions on Knowledge and Data Engineering* **18(1)** 63

[21] Freund Y, and Schapire RE 1997 *Journal of Computer and System Sciences* **55(1)** 119

[22] Sun Y, Kamel M, Wong A, Wang Y 2007 *Pattern Recognition* **40** 3358

[23] Schapire RE and Singer Y 1999 *Machine Learning* **37** 297

[24] Guo H and Viktor HL 2004 *SIGKDD Explorations* **6(1)** 30

[25] Chawla NV, Lazarevic A, Hall LO and Bowyer KW 2003 SMOTEBoost : Improving Liu X Y, Wu J, Zhou Z 2009 *IEEE Transactions on Systems, Man, and Cybernetics, Part B* **39(2)** 539

[26] Prediction of Minority Class in Boosting, *7th European Confrence on Principles and Practice Of Knowledge Discovery in Database* (Croatia: Dubrovnik) 107