Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 17, No. 4, December 2017, pp. 229-234
http://dx.doi.org/10.5391/IJFIS.2017.17.4.229

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions

Hansoo Lee, Jonggeun Kim, and Sungshin Kim
Department of Electrical and Computer Engineering, Pusan National University, Busan, Korea

Abstract

Sufficient amount of learning data is an essential condition to implement a classifier with excellent performance. However, the obtained data usually follow a significantly biased distribution of classes. It is called a class imbalance problem, which is one of the frequently occurred issues in the real world applications. This problem causes a considerable performance drop because most of the machine learning methods assume given data follow a balanced distribution of classes. The implemented classifier will derive false classification results if the problem is not solved. Therefore, this paper proposes a novel method, named as Gaussian-based SMOTE, to solve the problem by combining Gaussian distribution in a synthetic data generation process. It is confirmed that the proposed method could solve the class imbalance problem by conducting experiments with actual cases.

Keywords: Skewed class distribution, SMOTE, Gaussian random variable, Classification

1. Introduction

Several essential conditions can make a reliable and accurate classifier including a sufficient amount of data. However, the collected data in real world frequently follows a skewed class distribution. In other words, the number of majority class data dominates the minority. Considering that most of the classification methods assume equally distributed dataset, the skewed distribution can cause a significant performance loss. It is called as a class imbalance problem. If the feature space of given dataset is a high-dimension, the problem severely makes the performance of classifier worse. Therefore, the classifier has no choice but to produce unreliable and poor classification results without pre-processing [1].

A lot of strategies have been proposed to solve the issue [2–5]. The strategies can be categorized into three representative approaches: random sampling [6, 7], algorithmic modification [8], and cost-sensitive learning [9, 10]. Among them, the random sampling based methods are popular choices, which derive numerical balance between the majority and the minority: the former decreases the number of majority class data, while the latter increases the minority class data. They allow classifiers to be learned from the given data without bias. However, the classical random sampling method selects samples by using random sampling with replacement, which has less influence to improve the performance of classifiers.

Therefore, more sophisticated models have been proposed to deal with the problem as mentioned earlier. Of that, synthetic minority oversampling technique (SMOTE) [11] has become one of the most renowned solutions to resolve the class imbalance problem. It creates synthetic data based on the feature space similarities between existing minority samples.

Received: Dec. 8 2017
Revised : Dec. 18, 2017
Accepted: Dec. 19, 2017

Correspondence to: Sungshin Kim
(sskim@pusan.ac.kr)
©The Korean Institute of Intelligent Systems

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

In other words, the SMOTE method generates synthetic samples based on a combination of k-nearest neighbors method and uniform probability distribution until the required amount of samples are made. However, without considering neighborhood information of the minority class samples, the synthetic samples are in over-generalization problem, which is not helpful to resolve the class imbalance. Various adaptive sampling techniques have been proposed to solve the limitation including Borderline-SMOTE [12], adaptive synthetic sampling [13], safe-level-SMOTE [14], and SMOTE-RSB* [15].

In this paper, we focus on the issue that the synthetic samples tend to be generated on the line between the minority samples. If there is a significant gap between the majority and the minority, an enormous amount of synthetic data needs to be created. It means that the synthetic data tends to be placed on the same line with high probability. It can be considered as one of the types of over-generalization problem. The proposed method, named as Gaussian-based SMOTE, can solve the problem by combining Gaussian probability distribution in the feature space. The Gaussian probability distribution can make the SMOTE algorithm to generate new artificial samples deviated in the line but not significantly.

The rest of paper is organized as follows: at first, Section 2 gives brief explanations of the SMOTE algorithm. After that, Section 3 introduces the proposed Gaussian-based SMOTE. Section 4 provides experimental results and conclusion is given in Section 5.

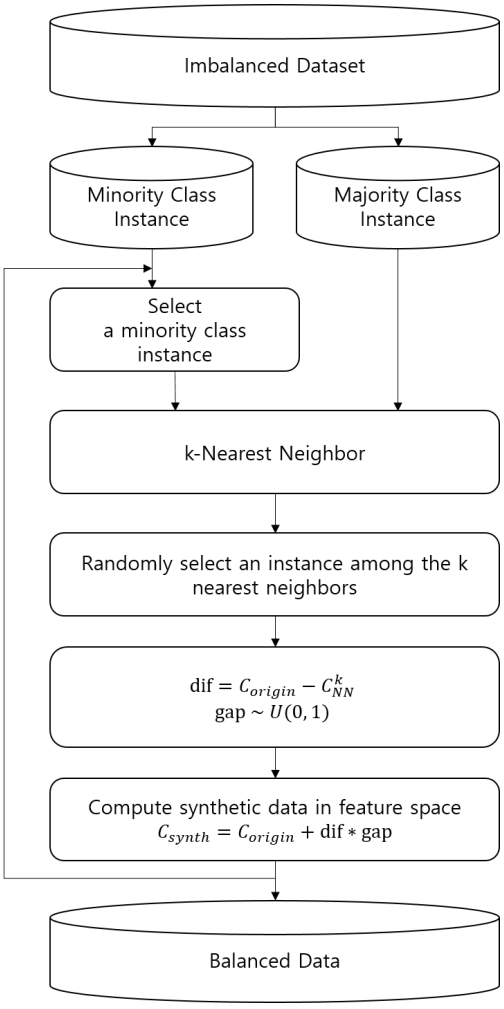

Figure 1. Flowchart of the SMOTE algorithm.

2. SMOTE

The SMOTE method generates synthetic data using k-nearest neighbor and uniform probability distribution. The operation principles are as follows. In the beginning, the method separates the given data into the majority and the minority class. After that, each minority class data will have the k number of its neighbors by the k-nearest neighbors algorithm. In the process of creating synthetic samples, each minority sample has its own randomly selected nearest neighbor among the k-nearest neighbors. Then find the remainder of the minority sample and its selected nearest neighbor as Eq. (1)

$$dif = \left| C_{origin} - C_{NN}^{k} \right|, \tag{1}$$

where C_{origin} is the minority class data and C_{NN}^{k} is one of k-th neighbors of the minority class data by random selection from uniform probability distribution. After that, the difference is multiplied by a random value from the uniform probability distribution to add randomness. Finally, it is possible to obtain a synthetic sample by Eq. (2).

$$C_{synth} = C_{origin} + \left| C_{origin} - C_{NN}^{k} \right| \times P_{uniform}, \tag{2}$$

where $P_{uniform}$ indicates a random value from the uniform probability distribution. The algorithm repeats the above sequences until satisfying stopping criterion including the number of generated samples. The entire process of the SMOTE algorithm is described in Figure 1.

3. Gaussian-Based SMOTE

The SMOTE, Borderline-SMOTE, and safe-level-SMOTE algorithms generate synthetic data by utilizing a random number from a uniform distribution. However, it is possible to hap-

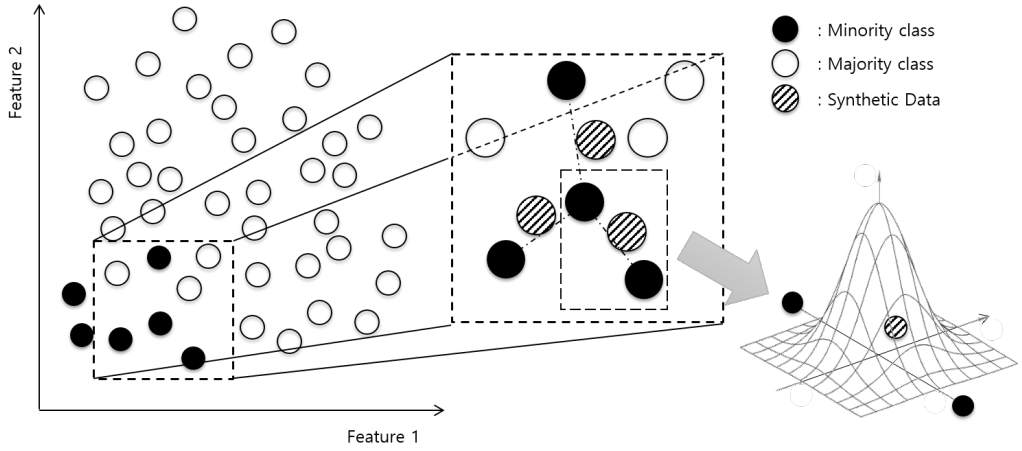

Figure 2. The operation principle of the Gaussian-based SMOTE algorithm.

pen that more than one synthetic data can be created between specific minority class data and its particular nearest neighbor, which are frequently selected during the process. In other words, the synthetic data are placed on the same line with high probability, which could intensify the over-generation problem.

Therefore, we propose a novel SMOTE algorithm, named as Gaussian-based SMOTE, to allow the method assure more diversity while generating artificial samples. The basic underlying principles of the proposed algorithm are same as the SMOTE method: compute differences between the minority class data and its randomly selected nearest neighbor in Eq. (1).

After that, the Gaussian-based SMOTE method draws a number between 0 and difference value for roughly estimating a location of a synthetic candidate as shown in Eq. (3).

$$\text{gap} \sim U(0, \text{dif}). \tag{3}$$

As a next step, another number draws from Gaussian (or normal) distribution as Eq. (4) with heuristically selected parameter, σ.

$$\text{range} \sim N(\text{gap}.\sigma) \tag{4}$$

Finally, a synthetic data is generated as Eq. (5) by using derived parameters in Eq. (3) and Eq. (4).

$$C_{synthetic} = C_{origin} + \text{dif} \times \text{range}. \tag{5}$$

The overview of Gaussian-based SMOTE is described in Figure 2. By including the Gaussian probability distribution in the process, it is possible to expand the place, where the synthetic sample is generated, from the line between minorities. Also, the Gaussian distribution allows the synthetic data located near the line. It makes the algorithm reasonable because the position of too far from the line might occurs false results. The deviated location between the minority classes provides the classifier a wealth of information.

4. Experimental Results

To verify the proposed Gaussian-based SMOTE method, we select a benchmark dataset first. Figure 3(a) indicates a seeds dataset from UCI Repository [16]. The seeds dataset consists of seven attributes and three classes. Each class is made of seventy samples. In this paper, we consider class 1 as the minority class, and others as the majority class. Before applying the SMOTE and the Gaussian-based SMOTE, we normalized the benchmark dataset from zero to one. As shown in Figure 3(a), the minority class samples are placed in the middle of the data distribution, and the majority class samples surrounding the minority class samples.

The synthetic sample generation results are described in Figure 3(b) and (c). Figure 3(b) is the result of using the SMOTE algorithm, and Figure 3(c) is the result of using the Gaussian-based SMOTE algorithm. As shown in the figures, the synthetic samples by the SMOTE algorithm seem to be generated in duplicated place. On the other hands, the synthetic samples by the Gaussian-based SMOTE algorithm seem to be of wide distribution nearby the minority samples. In conclusion, it is possible to consider that the proposed Gaussian-based SMOTE algorithm shows better performance than the SMOTE algorithm. To compare the performance numerically, we applied the support vector machine [17] as a classification method and comparing

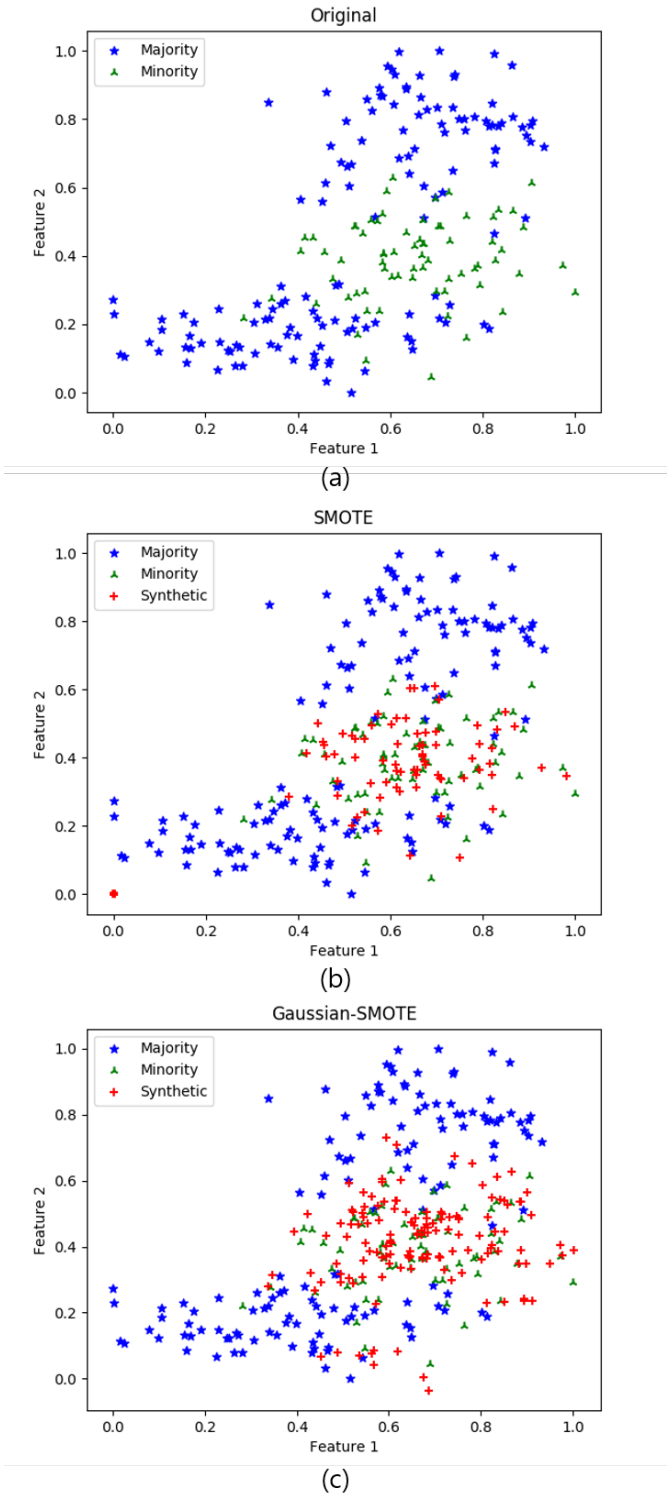

Figure 3. Benchmark dataset and synthetic samples: (a) given dataset, (b) SMOTE, (c) Gaussian-based SMOTE.

the performance by using the accuracy as shown in Eq. (6).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{6}$$

where TP stands for true positive, TN for true negative, FP for false positive and FN for false negative. Also, we conducted the experiments five times and derived the average performance for considering randomness in the algorithm. When we use the original dataset, the accuracy of the support vector machine shows 87.30% accuracy. And using the artificial dataset by the SMOTE, the accuracy shows 89.11%. However, when we use the Gaussian-SMOTE based synthetic dataset, the accuracy shows 90.13%.

Also, we select an interesting topic which is related to weather forecasting to verify the proposed method by utilizing a real world application. It is essential to remove non-weather echo in the radar data for obtaining high accuracy and reliability. An anomalous propagation echo is one of the non-weather echoes when a weather radar performs its observation process. It could cause significant false prediction results in quantitative precipitation estimation. Moreover, it occurs rare and random which is possible to consider as a class imbalance problem.

For comparison, we conducted experiments with the support vector machine using imbalanced class dataset and balanced dataset by the SMOTE and the Gaussian-based SMOTE. Because we implemented a binary classification method, we selected accuracy, sensitivity, and specificity as performance evaluations shown in Eq. (6), Eq. (7) and Eq. (8), respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{8}$$

An average accuracy using the Gaussian-based SMOTE is 87.45%, which is higher than 83.22% using imbalanced class dataset directly, and 85.66% using the SMOTE. Also, it turns out that the sensitivity and the specificity of the Gaussian-based SMOTE show 3% to 5% higher performance than others. Therefore, it is confirm that the proposed method could solve the class imbalance problem better than the SMOTE algorithm.

5. Conclusion

It is essential to provide sufficient amount of learning data in the implementation process of classifiers. However, learning samples from the real world include lots of problems including noise and skewed class distribution. The skewed class

distribution also called the class imbalance problem, causes a considerable performance loss due to the underlying assumption of machine learning methods. Therefore, it is important to resolve the problem of obtaining classifiers with excellent performance. In this paper, we proposed a novel method named Gaussian-based SMOTE by combining Gaussian probability distribution and the SMOTE algorithm.

It is confirmed that the proposed Gaussian-based SMOTE algorithm shows better performance than the SMOTE algorithm by using the benchmark data and the real-world application. In future work, we will deal with several considerations. At first, more performance comparison experiments should be conducted by utilizing other SMOTE algorithms including Borderline-SMOTE and safe-level-SMOTE. By doing this, it is possible to see which algorithm is better at various point of views such as accuracy, computational time, and start looking a new direction to improve the proposed Gaussian-SMOTE algorithm. Second, we will implement a method how to set the appropriate hyper-parameters including σ. Also, by conducting more experiments with other benchmark data, we will improve the performance to generate synthetic samples by combining other machine learning techniques such as clustering.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgement

This work was supported by the Energy Efficiency & Resources Core Technology Program of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20151110200040).

References

[1] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam: Elsevier, 2011.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009. http://doi.org/10.1109/TKDE.2008.239

[3] V. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113-141, 2013. https://doi.org/10.1016/j.ins.2013.07.007

[4] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1-6, 2004. http://doi.org/10.1145/1007730.1007733

[5] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429-449, 2002.

[6] N. V. Chawla, "Data mining for imbalanced datasets: an overview," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer, 2009, pp. 875-886.

[7] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20-29, 2004. http://doi.org/10.1145/1007730.1007735

[8] B. Zadrozny and C. Elkan, "Learning and making decisions when costs and probabilities are both unknown," in *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2001, pp. 204-213. http://doi.org/10.1145/502512.502540

[9] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proceedings of 3rd IEEE International Conference on Data Mining*, Melbourne, FL, 2003, pp. 435-442. http://doi.org/10.1109/ICDM.2003.1250950

[10] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007. https://doi.org/10.1016/j.patcog.2007.04.009

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[12] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets

learning,"in *Advances in Intelligent Computing*. Heidelberg: Springer, 2005, pp. 878-887. https://doi.org/10.1007/11538059_9

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of IEEE International Joint Conference on Neural Networks*, Hong Kong, 2008, pp. 1322-1328. http://doi.org/10.1109/IJCNN.2008.4633969

[14] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: safe-level-synthetic minority oversampling technique for handling the class imbalanced problem," in *Advances in Knowledge Discovery and Data Mining*. Heidelberg: Springer, 2009, pp. 475-482. https://doi.org/10.1007/978-3-642-01307-2_43

[15] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "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*, vol. 33, no. 2, pp. 245-265, 2012. https://doi.org/10.1007/s10115-011-0465-6

[16] M. Lichman, "UCI machine learning repository," 2013; Available http://archive.ics.uci.edu/ml/index.php

[17] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2001.

Hansoo Lee received his B.S. and M.S. degrees in Department of Electrical and Computer Engineering from Pusan National University, Korea, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree in Department of Electrical and Computer Engineering at Pusan National University, Korea. His present interests include intelligent system, machine learning, deep learning and complex data analysis.
Email: hansoo@pusan.ac.kr

Jonggeun Kim received his B.S. and M.S. degrees in Department of Electrical and Computer Engineering from Pusan National University, Korea, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in Department of Electrical and Computer Engineering at Pusan National University, Korea. His present interests include data mining, intelligent control, system modeling and fault diagnosis.
Email: wisekim@pusan.ac.kr

Sungshin Kim received his B.S. and M.S. degrees in Electrical Engineering from Yonsei University, Korea, in 1984 and 1986, respectively, and his Ph.D. degree in Electrical Engineering from the Georgia Institute of Technology, USA, in 1996. He is currently a professor at the Department of Electrical Engineering, Pusan National University. His research interests include fuzzy logic controls, neuro fuzzy systems, neural networks, robotics, signal analysis, and intelligent systems.
E-mail: sskim@pusan.ac.kr