|  |
| --- |
| ***Classification of Imbalanced Data*** |
| Sang-Hoon Oh\*  \*Mokwon University, Korea  E-mail:ohsanghoon16@gmail.com |
|  |

# Introduction

Since classifiers are usually developed under the assumption that the number of samples in each class is balanced with the others, it is difficult to accurately classify rare patterns/events and abnormal behavior. However, rare patterns/events and abnormal behavior result in heavy costs when we fail to figure out them in application problems. Rare patterns/events and abnormal behavior occur much less frequently than commonly occurring patterns/events and normal behavior. For example, software defects, natural disasters, cancer gene expression, fraud credit card transactions and telecommunications fraud belong to rare events or abnormal behavior. When one or some of classes have a much fewer number of examples than the others, we say “the dataset is imbalanced” and classifying the imbalanced data should have a different approach with classifying the balanced data. Rare patterns/events and abnormal behavior have the characteristics of the imbalanced data problem.

In this paper, we introduce the characteristics of imbalanced data problem and how to handle with the imbalanced data for better performance.

# What is Imbalanced Data Problem?

In classification applications, machine learning models are trained with data that have information related to the classification problems. If we use the Bayes classifier, the decision rule is to find the class with the maximum posterior probability. If we use the discriminant-based classifier, we attain the decision boundary of classes through training of data with class labels. In the case of balanced data problem, as shown in Fig. 1, the decision boundary will be a separator between classes. However, in the case of imbalanced data problem as shown in Fig. 2, the decision boundary will be severely distorted. As a result, the classification accuracy of the minority class will be heavily degraded. This is the imbalanced data problem.

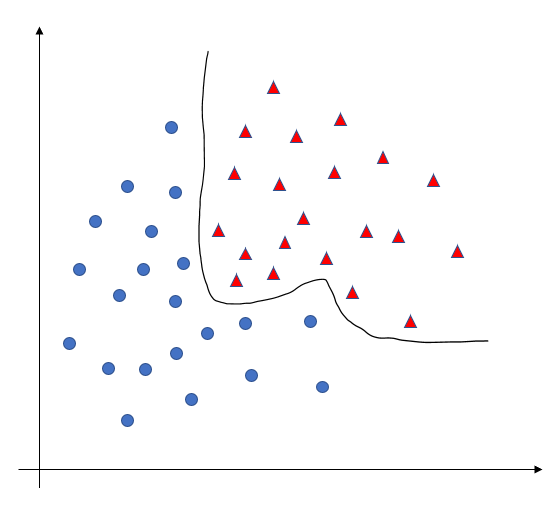


Fig. 1. Decision boundary with balanced data.

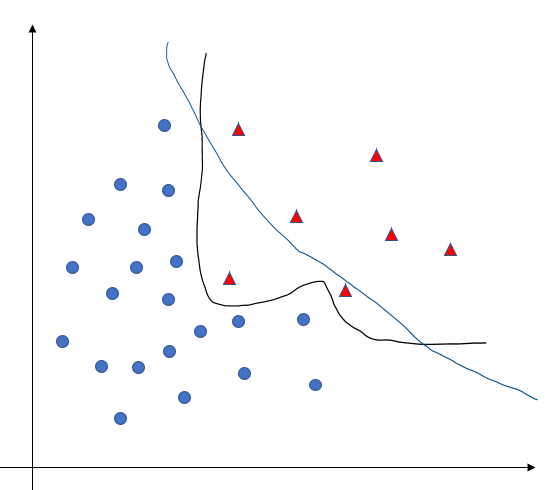


Fig. 2. Decision boundary with imbalanced data(Blue line).

# Methods for Imbalanced Data Problem

To resolve the imbalanced data problem, there have been many research results that can be categorized into data-level approach and algorithmic-level approach. Firstly, in the data-level approach, over-sampling is the technique to increase the number of samples in the minority class. Two widely-used methods are randomly duplicating the minority samples and creating minority samples using SMOTE[1]. Under-sampling is the technique to discard the intrinsic samples in the majority class[2]. Also, there is the hybrid technique which is the combination of the over-sampling and under-sampling technique. This data-level approach is popular in practical for laymen because of requiring no knowledge of machine learning models.

In the algorithmic approach, cost-sensitive learning is the method assuming higher costs for misclassification of samples in the minority class and strengthen the updating amounts of parameters related to the minority class[3]. In this case, cost matrix with elements represents the misclassification cost of assigning samples in class *i* to class *j*. However, it is difficult to set the values Because of the difficulty to set , there is a modified error function method to strengthen or weaken the updating amount of parameters regarding whether training samples are in the majority class or the minority class[4]. Also, there is an ensemble method that combines many classifiers using bagging or boosting[5].

# 4. Discussion and Conclusion

In this presentation, we briefly introduced the imbalanced data problem and the approaches to resolve the problem.

5. References

[1] N.V. Chawl et al., “SMOTE: Synthetic Minority Over-Sampling Technique” Journal of Artificial Intelligence Research, 2002, pp. 321-357.

[2] M. A. Tahir et al., “A Multiple Expert Approach to the Class Imbalance Problem Using Inverse Random Sampling,” Multiple Classifier Systems, 2009, pp. 82-91.

[3] V. Lopez et al., “Analysis of Preprocessing vs. Cost-Sensitive Learning for Imbalanced Classification,” Expert Systems with Applications, vo. 39, no. 7, 2012, pp. 6585-6608.

[4] S.-H. Oh, “Error Back-Propagation Algorithm for Classification of Imbalanced Data,” Neurocomputing, vol. 74, no. 6, 2011, pp.1058-1061.

[5] Z. Sun et al., “A Novel Ensemble Method for Classifying Imbalanced Data,” Pattern Recognition, vo. 48, no. 5, 2015, pp. 1623-1637.