Evaluating Oversampling Techniques for Financial Credit Risk Prediction: A Dataset Characteristic-Based Approach

[1] Yue Yang, [1] Boon Giin Lee, [1] Anthony Graham Bellotti, [1] Tangtangfang Fang, [1] Honghao Zhang [1] Junhan Xue

[1] School of Computer Science,
University of Nottingham Ningbo China
Ningbo, Zhejiang 315100, China

ABSTRACT

Financial credit risk prediction, e.g., credit card fraud detection, is a typical class imbalance problem where the sample size of the fraud transactions is relatively smaller. High false detection rate of credit card fraud can lead to huge financial losses. Hence, identification of fraud transaction (minority class) is important to reduce potential financial risks. Many existing studies apply oversampling techniques to address the class imbalanced issues when training with data hungry machine learning (ML) models. Nevertheless, the state-of-the-art oversampling techniques lack generalizability and are often tailored to a particular financial dataset only. Thus, this study aims to investigate the relationship between the financial data pattern in accordance with various state-of-the-art oversampling techniques. Considering this, our study proposes a method to evaluate the suitability of oversampling techniques in different financial dataset. The work starts by applying state-of-the-art oversampling methods on 15 financial benchmark datasets to generate synthetic samples and comparing the evaluation metrics of each augmented financial dataset with benchmark using statistical methods based on the sample size and data normality characteristics of the dataset. Then, categorize the datasets by analyzing their characteristics including class ratio, discreteness, and class distribution. The appropriateness of the proposed method of categorization is validated using different ML evaluation metrics such as precision, recall, F1-score, and AUC. The experimental results show that the oversampling methods have highly significant improvement of trained models on datasets with discreteness and non-time series characteristics/patterns (i.e., South German Credit dataset) compared to others (i.e., Australian Credit Approval dataset). In general, our work provides valuable insights in identifying the used of suitable oversampling method and machine learning model based on the characteristics of financial dataset, which could provide substantial economic impact on the financial market.