ABSTRACT

This study focuses on addressing class imbalance in data used as input for classification processes. Such imbalance is common in real-world scenarios—for instance, in customer churn prediction in the telecommunications or banking industries, where the number of churner samples (minority class) is significantly lower than that of non-churners (majority class). If not handled properly, the minority class may be overlooked, potentially leading to substantial revenue loss. Various sampling-based approaches have been proposed to address this issue at the data level. However, data quality problems, such as the presence of noise, can disrupt the sampling process and negatively affect classification performance, particularly in widely used oversampling techniques like the Synthetic Minority Over-sampling Technique (SMOTE). To address this, we propose an evolutionary hybrid sampling method to enhance classification performance on imbalanced data by improving data quality both before and after the SMOTE balancing process. The proposed method, called Tomek-SMOTE-GA (TSGA), is designed for binary-class problems. TSGA first applies sample denoising using the Tomek Links method before performing SMOTE, followed by sample optimization using an evolutionary algorithm. A Genetic Algorithm (GA), one of the most popular evolutionary algorithms, is employed for this optimization process. The resulting selected training set is then used to build classification models using five classifiers, i.e., Decision Tree, Logistic Regression, Support Vector Machine, k-Nearest Neighbors, and Naïve Bayes. Experimental results and statistical evaluations on 24 real-world imbalanced datasets demonstrate that the proposed method significantly better than both baseline and state-of-the-art sampling techniques in terms of geometric mean, particularly when using the Decision Tree classifier. These findings address two research questions posed in this dissertation. This study recommends further development of the TSGA method to address imbalanced problems in multi-class datasets, time-series data, and real-time application.

Keywords: imbalanced data, noise, SMOTE, evolutionary hybrid sampling, Tomek-SMOTE-GA (TSGA)