iBRF: Improved Balanced Random Forest Classifier

Abstract—

Class imbalance poses a major challenge in different classification tasks, which is a frequently occurring scenario in many real-world applications. Data resampling is considered to be the standard approach to address this issue. The goal of the technique is to balance the class distribution by generating new samples or eliminating samples from the data. A wide variety of sampling techniques have been proposed over the years to tackle this challenging problem. Sampling techniques can also be incorporated into the ensemble learning framework to obtain more generalized prediction performance. Balanced Random Forest (BRF), RUSBoost, and SMOTE-Bagging are some of the popular ensemble approaches used in imbalanced learning. In this study, we propose a modification to the BRF classifier to enhance the prediction performance. In the original algorithm, the Random Undersampling (RUS) technique was utilized to balance the bootstrap samples. However, randomly eliminating too many samples from the data leads to significant data loss, resulting in a major decline in performance. We propose to alleviate the scenario by incorporating a novel hybrid sampling approach to balance the uneven class distribution in each bootstrap sub-sample. Our proposed hybrid sampling technique, when incorporated into the framework of the Random Forest classifier, termed as ’iBRF: improved Balanced Random Forest classifier’, achieves better prediction performance than other sampling techniques used in imbalanced classification tasks. Experiments were carried out on 44 imbalanced datasets on which the original BRF classifier produced an average MCC score of 47.03% and an F1 score of 49.09%. Our proposed algorithm outperformed the approach by producing a far better MCC score of 53.04% and an F1 score of 55%. In addition, the algorithm is compared with 14 other benchmarking sampling techniques. Our proposed algorithm outperformed the other approaches by a large margin signifying its superiority and potential to be an effective sampling technique in imbalanced learning.