Comparison of imbalanced data treatments: a case study on cleft lip and palate data

ABSTRACT

This study was conducted to investigate if the resampling and the penalized approaches of balancing a small and imbalance data would improve the classification model produces by random forests learning algorithm on a small and imbalanced Cleft Lip and Palate (CLP) patients' dataset. Comparison between a Balanced Random Forest (BRF), Synthetic Minority Over-sampling Technique (SMOTE) on Random Forests (RF) and Weighted Random Forest (WRF) were then conducted on the CLP dataset and results were compared using the area under the curve (AUC) and the tradeoff between Sensitivity and Specificity. The results showed no difference in predictive ability between untreated (RF), oversampling (SMOTE+RF) and penalty treatment (WRF) but poor performances of the downsampling treatment (BRF). It was observed that the small number of training and test sample size had attributed to the results obtained and severely affect the performance of the classifier used for each treatment. The SMOTE+RF oversampling method, however, demonstrated to be promising for the CLP dataset.