Abstract

In this digital era, machine learning is a technology that is in demand by organizations and individuals. In the data and digital information age, the ability to process data efficiently is needed. As the amount of data grows, machine learning has various problems. One of them is that a class imbalance is also often found with the increasing amount of data. Class imbalance is a condition where a class dominates another class. One example is when the positive value class has fewer numbers than the negative class. The class that is less in number is categorized as the minority class, while the class that dominates the data set is called the majority class. Class imbalance can affect classification performance incorrectly, so handling imbalanced classes is needed to improve classification results. Classification of imbalanced data using Random Forest has satisfactory results, as well as implementing SMOTE and ADASYN as sampling methods because they are prevalent and easy to implement. In this research, we use ecoli protein data set to evaluate the performance of random forest classifier with and without oversampling methods. In this research, we used the f1 score and balanced accuracy as the primary evaluation metrics. We calculated the average with the highest score of 84% for the f1 score and 90% for balanced accuracy. Both SMOTE and ADASYN perform similarly to improve the classification performance and found that balanced accuracy is a better-suited metric for imbalanced classification.

Keywords: Imbalanced Data; Random Forest; Imbalanced Ratio; SMOTE;ADASYN