COLLEGE OF ENGINEERING AND COMPUTER SCIENCE

Comparison of Data Sampling Approaches for Imbalanced Bioinformatics Data

David Dittman, Randall Wald, Amri Napolitano and Taghi M. Khoshgoftaar, Ph.D.

College of Engineering and Computer Science, Florida Atlantic University

Class imbalance is a frequent problem found in bioinformatics datasets. Unfortunately, the minority class is usually also the class of interest. One of the methods to improve this situation is data sampling. There are a number of different data sampling methods, each with their own strengths and weaknesses, which makes choosing one a difficult prospect. In our work we compare three data sampling techniques Random Undersampling, Random Oversampling, and SMOTE on six bioinformatics datasets with varying levels of class imbalance. Additionally, we apply two different classifiers to the problem 5-NN and SVM, and use feature selection to reduce our datasets to 25 features prior to applying sampling. Our results show that there is very little difference between the data sampling techniques, although Random Undersampling is the most frequent top performing data sampling technique for both of our classifiers. We also performed statistical analysis which confirms that there is no statistical difference between the techniques. Therefore, our recommendation is to use Random Undersampling when choosing a data sampling technique, because it is less computationally expensive to implement than SMOTE and it also reduces the size of the dataset, which will improve subsequent computational costs without sacrificing classification performance.