Learning from Imbalanced Datasets: Evaluating the Predictive Accuracy of Minority Classes

Adithi Deborah Chakravarthy, PhD Student, Computer Science

Faculty Mentors: Qiuming Zhu & Zhengxin Chen, Computer Science

The class imbalance problem is an important machine learning challenge where the proportion of the majority class is much higher than the proportion of the minority class. Many real world applications involving health data, text mining and fraud detection learn from such coherently imbalanced datasets. Traditional machine learning algorithms are likely to produce good accuracy due to an obvious bias towards the majority class. Thus, accuracy as a measure of performance for imbalanced data is not very meaningful since the classifier has poor predictive accuracy over the minority class. While previous work has focused on using the Synthetic Minority Over-sampling Technique (SMOTE) technique to address this problem, in this research, we use the Random Over-Sampling Examples (ROSE) resampling method in different variances to evaluate five classifiers namely C5.0, K-Nearest Neighbors (KNN), Neural Nets, Random Forest and Support Vector Machines (SVM) in terms of other measures of performance against two highly imbalanced datasets. Preliminary results show that other measures of performance provide a more meaningful insight into the predictive accuracy of the minority class and can be used to build more optimal classifiers for coherently imbalanced data sets.

Keywords: Binary Classification, Imbalanced Data, Class Imbalance, Sampling, Performance Measures