Summary 1

Shaun Pritchard

Florida Atlantic University

CAP 6778

September -20-2021

M. Khoshgoftaar

This research seeks to address the issue of class imbalance. It was the first known paper that dealt with the issue of class imbalance. Through the experimentation and testing of unbalanced classified datasets. This paper demonstrated through the implementation of classification performance metrics followed by an analysis of variance (ANOVA) to further demonstrate statistical significance, determine the significance of even distribution in classification data and determine how many majority class examples would prove to be included within the training dataset when the minority class is rare.

In class imbalance, the issue arises with minority data classes. This implies that data with true positive features are classified unevenly and in a much lower frequency than the majority data class with true negative features creating a very uneven distribution. To note, In data mining, a classification model is trained on labeled datasets to classify unlabeled records into different categories. These classes become an uneven distribution creating the minority class and a majority class subsequently causing the imbalance issue.

This experiment examined the impact of class distribution on learning performance when training data has as few as five and as many as 40 minority class examples and implemented nearly one million classifiers underperformance measures, 11 learners, and 10 datasets.

When considering overall classification precision metrics under the (ROC) receiver operating characteristic curve (AUC) was implemented to evaluate the number of negative class examples and determine the class distribution and rare events to find the mean AUC value for each #P ( positive classes) and %P(class distribution).

Through implementing 11 learners and 10 datasets, statistics for AUC values for #P determined that the ratio of 35% ratio of 3:1 for three of the four values of #P (10,20,40) whereas, the ratio 25.0% for #P equal to 5 was determined to be the optimal percentage ratio of 2:1. Statistical significance was ensured by repeating the experiment 30 times and implementing 894,300 classifiers. This research proved optimal classifiers give rare data events of minority classes.