Classifying Commit Messages: A Case Study in Resampling Techniques
SeyedHamid Shekarforoush
Robert Green
Robert Dyer
Bowling Green State University

In practice, there are a variety of real-world datasets that have an imbalanced nature where one of two classes dominates the data. These datasets are generally difficult to classify using machine learning algorithms as the skewed nature of the data has a significant impact on the training process. In order to combat this difficulty, many methods of under sampling and over sampling have been proposed in order to generate comparable data sets that are more easily classifiable. This study applies multiple resampling techniques to a set of commit messages that have been extracted from multiple Github and Sourceforge projects in order to answer the question, "Do developers discuss design?" This dataset is highly imbalanced with less than 15% of all commit messages being classified as design commits. Results demonstrate that the combined use of resampling as coupled with various classifiers is effective in improving classification performance.

Classifiers:
- Random Forest (RF)
- Decision Tree (DT)
- Support Vector Classification (SVC)
- Linear SVC (LSVC)
- Bernoulli Naive Bayes (BNB)
- Nearest Centroid (NC)

Experiment 1
- Training set: Sourceforge & Github
- Testing set: Github & Sourceforge

Experiment 2
- Training set: Sourceforge / Github
- Testing set: Github / Sourceforge

Results
- True positive (TP)
- False negative (FN)
- False positive (FP)
- True negative (TN)

Recall (R): True positive rate or sensitivity
Specificity (S): True negative rate
Precision (P): Positive predictive value
Accuracy (A)
F1-score (F1): Harmonic mean of Precision and Recall
G Mean (G)