Summary 3-1 &3-2

Shaun Pritchard

Florida Atlantic University

CAP 6778

September -23-2021

M. Khoshgoftaar

Summary 3-1 - Building Useful Models from Imbalanced Data with Sampling and Boosting

By comparing sampling and boosting techniques and evaluating 16,000 classifiers, the research provided solutions for class imbalances arising from the traditional classification algorithms in the domain of software quality prediction. In conclusion, final results show that while using data sampling when training datasets results in better performance, boosting most often performs exceedingly. This research and data derived was implemented using Weka with C4.5/J48 and JRip/RIPPER C4.4 classifiers.

In the study, an example was used to compare the accuracy of imbalanced Software Quality Prediction models. In most cases, Software Quality Prediction models are built to distinguish between modules with faults. Despite not considering the ramifications for those modules not susceptible to a fault.

Essentially, if only a small percentage of the modules in a software project are prone to fault, then a model can still achieve high accuracy by classifying all modules as not being prone to fault. As a result, maintaining an incorrect classification would lead to inefficiency and imbalance. This is because classification algorithms do not consider positive or negative classes when maximizing accuracy.

The observations conducted in this research sought out to alleviate the problems noted in the example through implementing the associations with class imbalance using 5 data sampling techniques, a boosting ensemble meta-learners algorithm which was distributed to 5 software quality prediction datasets using 10-fold cross-validation, and distributed with 2 performance measures The KolmogorovSmirnov statistic (KS) and the area under the ROC curve

(AUC). These techniques were implemented with 2 classifying learner algorithms J48 and JRip in Weka.

Data sampling was used to overfit or underfit data by reducing and increasing the amounts of examples belonging to the majority class. This research used 5 data sampling techniques including both random and intelligent over and under sampling techniques. The five sampling techniques used were Random oversampling (ROS), Random under-sampling(RUS), Wilson's Editing (WE), Synthetic Minority Oversampling Technique (SMOTE), and Borderline-SMOTE (BSM)

Data Sampling techniques:

- Random oversampling (ROS), randomly removed examples from the minority class overfitting the data
- Random undersampling(RUS), Randomly replicating examples of the minority class underfitting data
- Wilson's Editing (WE), under-sampling that removes noise in data
- Synthetic Minority Oversampling Technique (SMOTE), creates oversampling with new minority class examples by extrapolating between existing minority class examples.
- Borderline-SMOTE (BSM) oversampling by only creating new examples based on minority class examples that lie near the decision border in feature space

The second technique implemented a well known boosting technique called Adaboost or adaptive boosting, which is an iterative cost-sensitive ensemble learning technique that converts weak classifiers into strong ones, and is designed to improve modifying the weight of training examples after each iteration to produce accuracy. Essentially, Adaboost takes all the

examples and assigns equal weights, generates the week hypothesis through the iteration calculating the associated error with the hypothesis, adjusting the new distribution of training weights with the misclassified examples, combining the outputs from the weak learners to create strong learners which eventually improves the prediction power of the model.

The results found that the data sampling methods, while effective, did not produce better results than Adaboost. With data sampling under the performance measure of AUC RUC performed the best. Data sampling techniques under KS performance metrics resulted in better output for RUS, SM, and BSM with the J48 classifier. Implantation of the JRIP classifier resulted in optimum results with the AUC performance metric in RUS and SM while RUS, SM, BSM performed better under KS metrics. JRip under AUC showed the best results in the C12 features with RUS at 0.8809.

Results found that the Adaptive boosting algorithm proved versatile and optimal under the performance measure AUC using the JRIP classifier with the highest instance of 0.8973. Boosting, which was not designed or considered to address the class imbalance issues(during the time this research was conducted) Was shown to perform more superiorly and consistently overall.

Summary 3-2 - Similarity Analysis of Feature Ranking Techniques on Imbalanced DNA Microarray Datasets

This research used real-world imbalanced data derived from DNA microarray technology. Allows a researcher to test samples to measure the expression levels of large numbers of genes simultaneously. This paper examines the application of diversity between ensemble feature selection and similarity measure ranking techniques on high dimensional Imbalanced data. This study performed eighteen feature selection techniques (eleven of which were very similar), nine imbalanced datasets, and four feature subset sizes for a total of 162 different rankings.

The reason for choosing the imbalanced datasets is that even though the minority classes are the main interest, DNA microarray experiments often have very few samples or instances of the minority classes. Inturn using feature selection methods with an optimal subset of features could be processed and selected for use in later analysis, rather than analyzing the entire dataset at once. As a result, four subset sizes per rank of 50, 75, 100, and 200 were selected.

This experiment implemented three categories for the feature selection techniques during analysis: The common statistical filters(Chi-Squared, Information Gain, Gain Ratio, Symmetric Uncertainty, ReliefF, ReliefF-W), Signal to Noise (S2N), and Threshold-Based Feature Selection Techniques. The common filters and Signal to Noise filters are classified as non-TBFS, while Threshold-Based Feature Selection refers to TBFS. Note that TBFS features were normalized and evaluated independently against the class attributes. TBFS techniques consisted of eleven techniques (F-Measure, Odds Ratio, Power, Probability Ratio, Gini Index, Mutual

Information, Kolmogorov-Smirnov statistic, Deviance, Geometric Mean, Area Under the ROC Curve, and Area Under the Precision-Recall Curve)

The Similarity Measure calculates the intersection between the subsets, given a data set with n features, based on a consistency index, and determining the feature subset size of both subsets for comparison and similarity to find diversity. The comparable properties of multiple feature subsets were computed using this framework for examining the diversity of feature selection methods.

If the techniques are diverse, then there is more reason to have confidence in any genes found by all methods, while it is unsurprising (and not confidence-enhancing) when similar techniques all select the same genes.

It was implemented that diversity is an important factor when using ensemble feature selection because it allows for smaller ensembles that produce maximum performance in the classification. These are the diverse techniques used together to provide the final results. The analysis of eighteen filter-based feature selection techniques across nine imbalanced datasets. Certain clusters of feature selection techniques that exhibit high similarity could not be combined to ensure maximum diversity. The following, {PR and GI} and {KS and GM}. The {PR and GI} cluster is particularly dissimilar to all other rankers, and the {KS and GM} cluster shows the highest similarity for any pair of rankers. In addition, the {CS, IG, and SU} and {CS, Dev, F, MI} clusters showed high within-cluster similarity. overall, positive proving to result in a derivation that will allow for more specific cases of feature selection and similar further validating this type of high dimensional data.