**Summary 2 - Experimental Perspectives on Learning from Imbalanced Data**

Shaun Pritchard

Florida Atlantic University

CAP 6778

September -21-2021

M. Khoshgoftaar

**Summary 2 - Experimental Perspectives on Learning from Imbalanced Data**

In this study, a comprehensive and systematic experimental analysis was performed concerning learning from imbalanced data and improving the performance of classifiers in minority settings based on the combination of sampling techniques on multiple learning classifiers. The research shows how critical sampling is to improving classification performance especially optimizing threshold-dependent measures such as the geometric mean, TPR, and its contrast to cost-sensitive learning. Additionally, the data shows that decision tree learners have class imbalances that are unambiguous with other learning techniques such as neural networks, Logistic Regressions, and K-nearest Neighbor algorithms.

Researchers employed seven sampling techniques and 11 different learning algorithms with 35 benchmark datasets from a variety of applications to analyze the results of their experiments. Also, Null datasets and transformations requiring 31 test sets were developed for validation. The researchers used these results to prove that constructing optimized classifiers from imbalanced data could deliver accurate results and that threshold-dependent measures could be optimized, with a total of 1,232,000 classifiers constructed with 95% statistical confidence, and ANOVA analysis was performed using the SAS GLM procedure.

The research shows that data used was based on the percentage of minority examples from 1.33% (highly imbalanced) to almost 35% (slightly imbalanced). The 7 sampling techniques that were implemented were random undersampling (RUS), random oversampling (ROS), one-sided selection (OSS), cluster-based oversampling (CBOS), Wilson’s editing (WE), Synthetic Minority Oversampling Technique SMOTE (SM), and borderline-SMOTE (BSM).

The 7 sampling techniques were used in conjunction with 11 learner classification algorithms then being applied across 35 individual datasets for analysis. The experiment was conducted using Java/Weka. Note, default parameters were only changed when the inference showed improvement.

The 11 classifier algorithms implemented in Weka were as follows:

* 2 different versions of the C4.5/J48
* 2 KNN k-nearest neighbors classifiers (using k= 2, and k = 5 with distance Weighting parameter set to ‘Weight by 1/distance)
* Naïve Bayes classifier
* Multilayer perceptron's (MLP)
* Radial basis function networks
* RIPPER (Repeated Incremental Pruning to Produce Error Reduction)
* logistic Regression (LR)
* Random Forest (RF ensemble learning technique)
* Support vector machine (SVM)

Twenty different five-fold cross-validation (CV) runs were performed for each dataset. For the 4 folds there was a transformation of instances, while the last fold took care of applying the testing. Following the sampling techniques, 11 different classification learner's were constructed from the transform data set and tested.

This was Then followed by multiple classification performance measurements such as the area under the ROC curve (AUC), and Kolmogorov-Smirnov statistic (K/S). Then performance measurements of geometric mean (G), F-measure (F), accuracy (ACC), and true positive rate (TPR) were implemented and calculated using the probability of positive class membership of > 0.5.

The results of these studies facilitated individual learners, as the datasets were grouped into four categories depending on the degree of imbalance. π being the percentage of the minority class displaying the results of the classification and sampling class membership π < 5%, 5% <π< 10%, 10% <π< 20%, and finally 20% < π.

There were only 5 learners and only 2 performance measurements (AUC and G) presented due to space constraints in the tables generated for the inferences of the research document. The results were various but indicative of comparing learners, sampling techniques, and performance measures show that different types of sampling work best with different types of classifications of learners.

For example some sampling did not significantly improve the AUC obtained by NB. However applying either RUS, SM or BSM does significantly improve G. The data showed that RUS or ROS often performed much better. Using the AUC, sampling significantly improved upon the performance of the classifier constructed with the unaltered data in only 15 of 44 scenarios. Also, one of the most important conclusions was noted was the high performance of the intelligence sampling techniques such as SM, BSM, WE, OSS, and CBOS.

The research shows that RUS works better for C4.5D or RF, While ROS works significantly better with LR. Threshold dependency was a major factor that showed independent thresholds such as with AUC and K/S which generated different negative results. It was also determined that the value of sampling is dependent on the performance measure being used. For numerous learners, such as NB, LR, 2NN, 5NN, and none of the sampling techniques improved the performance of the learner measured by the AUC or K/S.

By comparing both sampling techniques and classification learners, the experiment demonstrated the efficacy of learners. The study meets its demand and provides an in-depth look at sampling techniques and the overall viability sampling has with performance of learners.