Summary 13

Shaun Pritchard

Florida Atlantic University

CAP 6778

November 2, 0-2021

M. Khoshgoftaar

The Effects of Class Label Noise on Highly-Imbalanced Big Data

This study implements simulated class label noise which is injected into big data sets for evaluating classification models. This study evaluated the effects of injecting the class label noise using over 8 million instances with highly imbalanced big data with an imbalance rate compared to 0.0975%. Six rounds of five-fold cross-validation were used to simulate the models in this project using four machine learning models. Metrics for measuring performance of the experiment implemented the use of Model classification performance used as the main performance metric for the study.

The dataset consists of highly imbalanced Medicare part B data from 2012 to 2018 pre-processed into a binary fraud detection dataset containing medical procedures performed by Medicare providers. For Medicare part B insurance claims classification, data was classified as non-fraudulent and fraudulent.

A noise ranking approach was used to clean the data set, comparing the 4118 positives to the 8443.052 negatives of the original dataset and the 4118 positives to the 4221.526 negatives of the cleaned dataset, we found -216 variance between them. The four learners that were evaluated and implemented with fivefold cross-validation 6 times each were Random Forest(RF), Logistic Regression(LR), Multilayer Perceptron(MLP), and XGBoost(XGB).

To calculate the true positive and negative rates, the Area Under the Precision Recall (AUPR) curve was used as the performance metric. These noise levels are simulated by applying Lambda and Psi to instances of class noise. Each model shows a significant reduction in classification performance across all noise levels and all learners. All models showed low AUPRC by significant margins. The results showed that the RF and MLP learners had similar AUPRC and

XGBoost had the highest AUPRC. In regards to the TPR and TNR, the XGBoost results show that its robustness to class label noise would be improved by reducing the complexity of the architecture.