Summary 15

Shaun Pritchard

Florida Atlantic University

CAP 6778

November 8, 0-2021

M. Khoshgoftaar

Robust Thresholding Strategies for Highly Imbalanced and Noisy Data

Using Part B Medicare data with 4.2 million samples and a positive class size of just 0.097%, this study evaluated the implementation of Thresholding Strategies for highly imbalanced binary classification problems. In the paper, it is shown that when training data are highly imbalanced, and using the positive class makes up * 1% of the data, learners will over classify the majority group and miss the minority group. During the experiment, the threshold lambda was set equal to the prior probability of the study of the positive class, and lambda was optimized using training data itself to test significance.

The research implements four thresholding strategies, two thresholds that are optimized based on training data and two thresholds based on the positive class prior. Using Random Forest, Multilayer Perceptron, and XGBoost classification learners, threshold strategies are evaluated for a range of noise levels and noise distributions. To determine the optimal performance thresholds, this study made use of G-Mean and F-Measure metrics and tested whether the proposed methods would exceed the default thresholds, and Under the Receiver Operating Characteristics Curve (AUC). Additionally, this was one of the first studies to consider implementing output thresholding for imbalanced and noisy big data sets.

To start with, they removed pre-existing label noise from the Medicaid part B data sets Dc & D (parameters to inject the noise). Next, they compared the performance of these thresholding strategies on noisy data sets, including the G-mean, F-measure, and prior thresholds equal to the positive class before and after noise injection.

Ultimately, it was found that three of the four thresholds are unstable in the presence of class noise when compared to the Geometric Mean (G-Mean). The results also show that

setting the threshold equal to the prior probability of the noisy positive class consistently results in the best performance according to G-Mean, TPR, and TNR. This study concluded that a threshold equal to the prior probability of the positive class lambad_np from a noisy distribution Dn consistently outperforms other thresholds, noise levels, and noise distributions.