Robust Thresholding Strategies for Highly Imbalanced and Noisy Data

Justin M. Johnson and Taghi M. Khoshgoftaar
College of Engineering and Computer Science
Florida Atlantic University
Boca Raton, Florida 33431
jjohn273@fau.edu, khoshgof@fau.edu

***Abstract*—Many studies have shown that non-default decision thresholds are required to maximize classification performance on highly imbalanced data sets. Thresholding strategies include using a threshold equal to the prior probability of the positive class or identifying an optimal threshold on training data. It is not clear, however, how these thresholding strategies will generalize to imbalanced data sets that contain class label noise. When class noise is present, the positive class prior is influenced by the class label noise, and a threshold that is optimized on noisy training data may not generalize to test data. We employ four thresholding strategies: two thresholds that are optimized on training data and two thresholds that depend on the positive class prior. Threshold strategies are evaluated on a range of noise levels and noise distributions using the Random Forest, Multilayer Perceptron, and XGBoost learners. While all four thresholding strategies significantly outperform the default threshold with respect to the Geometric Mean (G-Mean), three of the four thresholds yield unstable true positive rates (TPR) and true negative rates (TNR) in the presence of class noise. Results show that setting the threshold equal to the prior probability of the noisy positive class consistently performs best according to G-Mean, TPR, and TNR. This is the first evaluation of thresholding strategies for imbalanced and noisy data, to the best of our knowledge, and our results contradict related works that have suggested optimizing thresholds on training data as the best approach.**

***Keywords*—Class Noise, Class Imbalance, Output Thresholding, Big Data, Medicare, Fraud Detection**

1. Introduction

Class imbalance exists when the total number of samples from one class, or category, is significantly larger than any other category within the data set. This class-imbalanced data problem arises in many real-world applications [1], [2], [3], [4]. In each of these examples, the positive class of interest is usually the smaller class, i.e. the minority group, and there is an abundance of less-interesting negative samples comprising the majority group. When training data is highly imbalanced, and the positive class is $< 1\%$ of the data set, learners will typically over-classify the majority group and fail to detect the under-represented minority group [5]. In this study we focus specifically on the highly imbalanced binary classification problem. The techniques in this study can be extended to the multi-class problem, as multi-class problems can be converted into a set of two-class problems through class decomposition [6].

In the binary classification problem, we have data set $\mathcal{D} : \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$ where x_i and y_i correspond to the i^{th} feature vector and class label, respectively. A probabilistic classifier $\mathcal{M}$ that is trained on $\mathcal{D}$ is used to estimate the posterior probability $\mathcal{P}(y_i = 1 \mid x_i)$. The predicted class label is then determined by comparing the probability estimate to the decision threshold λ, i.e. $\hat{y}_i = 1$ if $\mathcal{P}(y_i = 1 \mid x_i) > \lambda$, else $\hat{y}_i = 0$. Generally, the default threshold of $\lambda = 0.5$ is used to assign class labels to test samples. Studies have shown, however, that the classification performance of highly imbalanced data sets can be improved significantly by using non-default thresholds [7], [8], [9]. Instead, these related works suggest setting the threshold equal to the prior probability of the positive class, i.e. $\mathcal{P}(y = 1)$, or optimizing the decision threshold using training data. It is not clear, however, how these thresholding strategies will perform when training data also exhibits class label noise. We believe this imbalanced and noisy data problem is an important consideration that is worth exploring further, as class label noise is common among imbalanced classification problems [10].

Class label noise occurs when the observed class label $\tilde{y}_i$ is incorrectly assigned a noisy label value, and the noise level $\mathcal{P}(\tilde{y}_i \neq y_i)$ denotes the probability x_i is mislabeled. In this paper, we refer to positive samples with corrupt negative labels as $P \rightarrow N$ noise and negative samples with corrupt positive labels as $N \rightarrow P$ noise. These types of erroneous labels can be the result of human error, subjective labeling tasks, non-exact data labeling processes, or malfunctioning data collection infrastructure [11]. This class noise will alter the training data's positive class prior and has been shown to negatively influence the learner's training and inference processes, especially when noise levels are high or heavily skewed towards one class. In this study we explore a range of noise levels and noise distributions to demonstrate the effect of class noise on thresholding strategies for classifying highly imbalanced data.

We evaluate thresholding strategies using a highly imbalanced Part B Medicare data set provided by the Centers

for Medicaid and Medicare Services (CMS) [12]. First, we create a clean subset of Part B data $\mathcal{D}^c \in \mathcal{D}$ by removing pre-existing label noise, yielding a big data set with 4.2 million samples and a positive class size of just 0.097%, i.e. 4118 positive fraudulent samples. Next, we create noisy training sets $\mathcal{D}^n$ using two class label noise parameters. The class noise level Λ determines the percentage of training samples whose class labels are corrupted. The noise distribution Ψ determines the proportion of class noise that is injected into the positive and negative class. Random Forest (RF), Multilayer Perceptron (MLP), and XGBoost [13] learners are trained on each $\mathcal{D}^n$ and scored on the clean test partitions using four different thresholding strategies. Overall performance is measured using the Geometric Mean (G-Mean), and the tradeoffs between positive and negative class performance is measured using the True Positive Rate (TPR) and True Negative Rate (TNR). Six rounds of five-fold cross-validation are used to compare the thresholding strategies. Due to space limitations, we restrict this study to the Part B data set with a wide range of noise distributions and leave the evaluation of additional data sets to future works.

The four thresholding strategies compared are the: 1) prior threshold, 2) noise-prior threshold, 3) optimal-fmeasure threshold, and 4) optimal-gmean threshold. The prior and noise-prior thresholds set the decision threshold equal to the prior probability of the positive class before noise injection and after noise injection, respectively. The optimal-fmeasure and optimal-gmean thresholds optimize decision thresholds on the training data using each respective performance metric. Results show that the prior, optimal-fmeasure, and optimal-gmean thresholds are generally unstable across various noise levels and distributions. In particular, these thresholds see large tradeoffs in TPR and TNR as noise levels change, leading to skewed classifications on the test set that favor one class over the other. The noise-prior threshold is shown to consistently outperform these three thresholds according to the G-Mean. Furthermore, TPR and TNR results show that the noise-prior threshold obtains approximately balanced results across all learners. Unlike related works that do not consider the class noise problem when selecting decision thresholds, our results show that the noise-prior threshold is the preferred thresholding strategy when class noise exists.

Section 2 introduces related works in the areas of output thresholding and classifying imbalanced data. Section 3 describes the data pre-processing, noise injection process, experiment design, and performance evaluation used in this study. Section 4 presents the results of our experiments and discusses key findings. Finally, Section 5 concludes with a summary of our results and suggestions for future works.

2. Related Work

Output thresholding has been widely used to improve the classification performance of imbalanced data sets. Buda et al. [7] applied the prior probability thresholding strategy to image classification problems using Convolutional Neural Networks. Results showed significant improvements to classification performance, especially when thresholding is combined with random over-sampling of the minority classes. Zou et al [14] identify optimal thresholds for imbalanced protein sequence classification by maximizing the F-Measure on the training set, and then tuning the threshold using the minimum and maximum probability estimates from the test set. The authors show that the proposed threshold strategy outperforms the default threshold and other uniform thresholds, and propose a method for scaling the thresholding selection process to big data. Xingfu et al. [9] propose a thresholding strategy that optimizes the threshold on the test set by selecting the threshold that minimizes the difference between the positive class frequency in the training set and the positive class frequency in the test set predictions. They extend this approach to the multi-label classification problem and show that it outperforms the default threshold. Calvert and Khoshgoftaar [15] compute optimal-gmean and optimal-fmeasure thresholds to improve classification performance on an imbalanced network security traffic data set. The authors use the TPR, TNR, and G-Mean metrics to show that the optimal thresholds significantly outperform the default threshold, and demonstrate that the popular Area Under the Receiver Operating Characteristics Curve (AUC) alone is insufficient for selecting the best model. Johnson and Khoshgoftaar explored the effects of output thresholding using MLP learners [8] and ensemble learners [16]. Results show that thresholds optimized on training data and prior thresholds consistently outperform the default threshold of 0.5 when the training data's classes are imbalanced. Each of these works support using non-default decision thresholds for classifying imbalanced data, but they do not consider how class noise will affect the threshold selection process. We compare how these thresholding strategies perform on noisy data sets, including the optimal-gmean, optimal-fmeasure, and prior thresholds equal to the positive class prior before and after noise injection. We do not compare results with those methods that optimize thresholds on the test set, as tuning parameters on test data is generally discouraged and doesn't lend itself to real-world machine learning applications having unknown test data [17].

The effects of class label noise, and methods for learning from noisy data, have been studied extensively by the machine learning community. Frénay and Verleysen [18] provide a comprehensive survey on classification in the presence of label noise. Topics covered include introductory definitions for label noise, its sources, a taxonomy for categorizing different types of label noise, and popular techniques for treating class noise. Techniques for addressing class label noise are broadly divided into two categories: data-level techniques and algorithm-level techniques. Data-level techniques, or noise filters, pre-process data to identify and discard or correct class noise. Distance-based filters [19], [20], ensemble classifier filters [21], [22], [23], and variants of these [24], [25], [26], [27] are among the most popular data-level methods. Algorithm-level techniques employ noise-robust [11], [28] or noise-tolerant classifiers that can be trained directly using noisy data. Noise tolerant classifiers

TABLE 1. PART B MEDICARE DATA SET SUMMARY

Part B Data Set	# of Positive	# of Negative	Total	% Positive
Original	4334	8443052	8447386	0.0513
Cleaned	4118	4221526	4225644	0.0975

TABLE 2. CORRUPTED SAMPLES FOR NOISE DISTRIBUTIONS ($P \rightarrow N$, $N \rightarrow P$)

	$\Psi = 0.0$	$\Psi = 0.25$	$\Psi = 0.5$	$\Psi = 0.75$	$\Psi = 1.0$
$\Lambda = 0.0$	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\Lambda = 0.1$	(0, 824)	(206, 618)	(412, 412)	(618, 206)	(824, 0)
$\Lambda = 0.2$	(0, 1647)	(412, 1235)	(824, 823)	(1235, 412)	(1647, 0)
$\Lambda = 0.3$	(0, 2471)	(618, 1853)	(1235, 1236)	(1853, 618)	(2471, 0)
$\Lambda = 0.4$	(0, 3294)	(824, 2470)	(1647, 1647)	(2471, 823)	(3294, 0)
$\Lambda = 0.5$	(0, 4118)	(1030, 3088)	(2059, 2059)	(3088, 1030)	(4118, 0)

include enhanced classifiers that down weight the impact of noisy samples and frameworks that use noise distribution estimates to alter the training and inference processes [29], [30], [31]. While research in this area is extensive, there is very little work that addresses both class label noise and class imbalance.

Van Hulse and Khoshgoftaar [10] explore techniques for learning from imbalanced and noisy data that considers the effect of the overall noise level (Λ) and the distribution of class noise (Ψ) across seven data sets and 11 classifiers. Results consistently show that the noise distribution Ψ, i.e. the proportion of corrupted labels in the minority and majority class, has the greatest effect on AUC performance. Overall, performance degradation is greatest when there are high levels of $P \rightarrow N$ noise. Several sampling methods are explored to address the class imbalance, and results show that under-sampling of the majority class can significantly improve performance across most learners. Kennedy et al. [32] perform similar noise experiments using highly imbalanced big data and show that regularization can be used to improve performance. In our study, we use the noise injection process proposed by Van Hulse and Khoshgoftaar, and expand upon their study by exploring how thresholding strategies can be used to improve classification performance across various class noise scenarios.

3. Methodology

3.1. Data Preprocessing

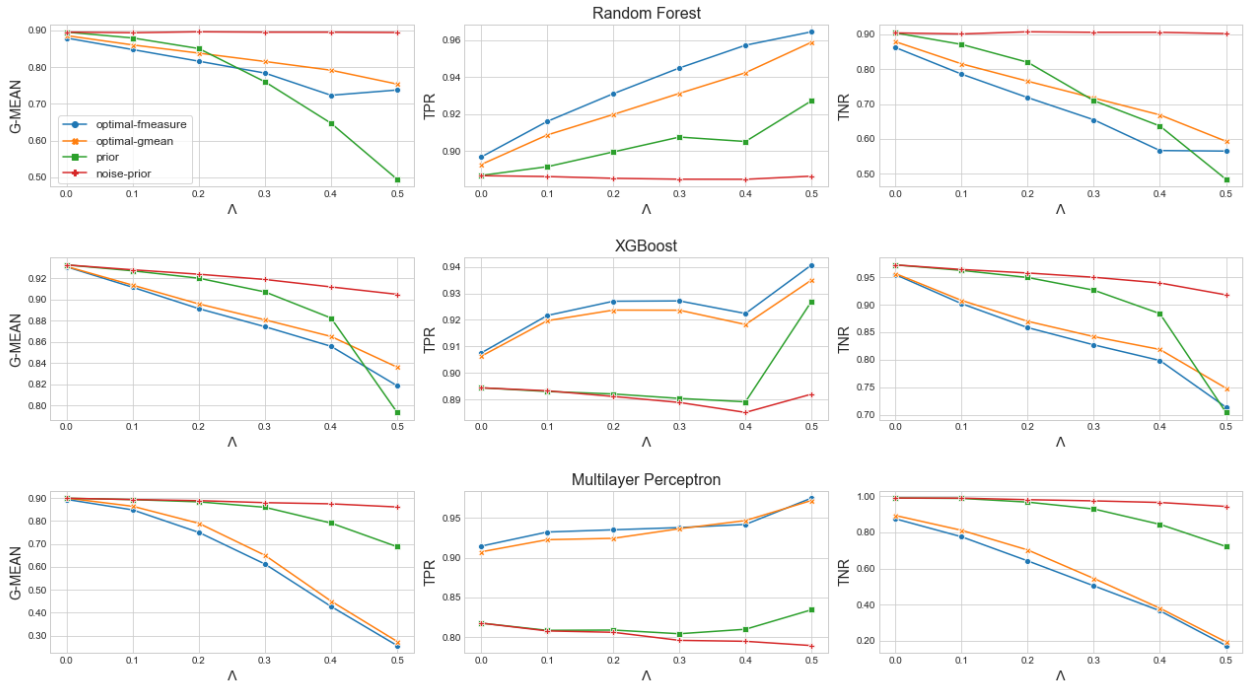
This study uses the 2012–2018 Part B Medicare data sets that have been made publicly available by the CMS [12]. The Part B data set is a highly imbalanced heterogeneous data set that describes the services and procedures performed by Medicare professionals. Real-world fraudulent labels are assigned to the Part B data set by joining with the List of Excluded Individuals and Entities (LEIE) [33]. Providers listed in the LEIE data set are prohibited from participating in the Medicare program for fraud-related offenses. Due to space limitations, we refer readers to a related work for a summary of features and the steps required to prepare the Medicare Part B data set for classification [34].

To properly evaluate the effect of the noise injection process, we first clean the Part B data set of pre-existing class noise using an ensemble classifier noise ranking technique. We create the clean subset $\mathcal{D}^c \in \mathcal{D}$ by training a RF classifier on $\mathcal{D}$ using five-fold cross validation and ranking the class-wise probability estimates from hold-out partitions. We remove the $K\%$ of samples with the lowest probability estimates from each respective class, and repeat this process for $K \in [0.0, 0.5]$. The final values $K_{pos} = 0.05$ and $K_{neg} = 0.5$ are selected by identifying the minimum values of K that obtain an AUC ≥ 0.95 when scored with a new RF classifier. In other words, we removed 5% of the positive samples and 50% of the negative samples to obtain a relatively clean and learnable Part B data set for noise experiments. A summary of the original Part B data set, the cleaned Part B data set, and the class imbalance levels are listed in Table 1.

3.2. Noise Injection

We use two parameters to inject noise into the training partitions of $\mathcal{D}^c$ to create noisy data sets $\mathcal{D}^n$ that simulate the various types of noise that exist in real-world applications [10]. The noise level Λ determines the total percentage of noise that is injected into the positive and negative classes of the training set. More specifically, the total number of corrupted samples is given by $2 \times \Lambda \times N_p$, where N_p is the total number of positive samples. Given the Part B data set $\mathcal{D}^c$ and $\Lambda = 0.2$, the total number of corrupted samples is $2 \times 0.2 \times 4,118 = 1,647$. The noise distribution Ψ determines the percentage of samples from the positive class that are corrupted. Following the same example, when $\Psi = 0.0$ then the total number of $P \rightarrow N$ corrupted samples is 0 and the total number of $N \rightarrow P$ corrupted samples is 1,647. Similarly, when $\Psi = 0.2$ then the total number of $P \rightarrow N$ corrupted samples is 239 and the total number of $N \rightarrow P$ corrupted samples is 1,318. We use $\Lambda \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\Psi \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$ to simulate a wide range of noisy data scenarios, as outlined in Table 2.

Figure 1. Classification Performance For Noise Levels Λ



3.3. Thresholding Strategies

The prior threshold (λ_p) and noise-prior threshold (λ_{np}) strategies are defined by the prior probability of the positive class in $\mathcal{D}^c$ and $\mathcal{D}^n$, respectively. The prior threshold is constant across all experiments, i.e. $\lambda_p = 0.00098$, and the noise-prior threshold changes for each noise distribution because it depends on Λ and Ψ. For example, when $\Lambda = 0.5$ and $\Psi = 0.0$, the positive class increases by 4,118 samples and the negative class decreases by 4,118 samples. This increases the prior probability of the positive class in the training data and yields $\lambda_{np} = 0.00195$.

The next two thresholding strategies are optimized on the training data after model training is complete. The optimal-gmean threshold (λ_{gmean}) is the threshold that maximizes the G-Mean metric on the training set. Similarly, the optimal-fmeasure threshold ($\lambda_{fmeasure}$) is the threshold that maximizes the F-Measure metric on the training set. Optimizing the threshold by the G-Mean metric seeks to approximately balance the TPR and TNR, while the F-Measure metric aims to maximize the harmonic mean of precision and recall. We add an additional constraint to both of these optimization processes, and require that $TPR \geq TNR$, as we are mostly concerned with detecting fraudulent providers. Additional details of these optimization processes can be found in related works [8].

3.4. Performance Evaluation

Learners $\mathcal{M}$ and thresholding strategies λ are evaluated across all noise distributions (Λ, Ψ) using five-fold cross-validation. For each iteration of cross-validation, a noisy training set $\mathcal{D}^n$ is created by corrupting the labels of the training partitions. Models are trained on $\mathcal{D}^n$ and scored on the clean hold-out partitions using each thresholding strategy. Each experiment is repeated six times to obtain 30 results for each $\{\mathcal{M}, \lambda, \Lambda, \Psi\}$. The G-Mean is recorded to capture the overall classification performance. TPR and TNR scores are recorded to identify any tradeoffs between class-wise classification performance. For the fraud detection application, we would like to maximize the TPR while maintaining an approximately balanced TPR and TNR.

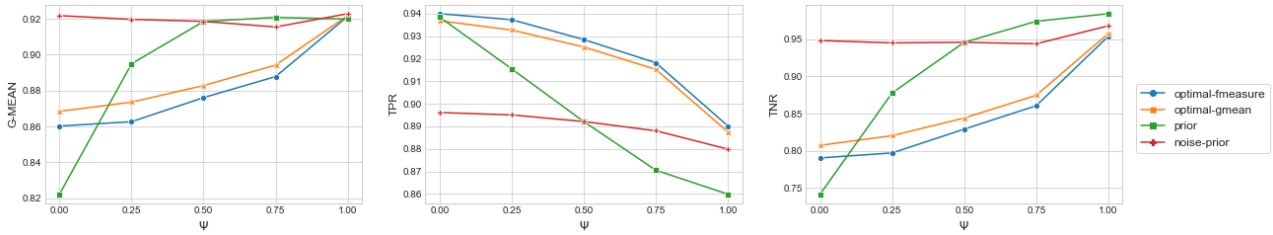
4. Results

First, we review the effect of increasing noise levels Λ on classification performance across all learners and threshold strategies using Figure 1. Next, we take the best performing learner (XGBoost) and explore how changes to the noise distribution Ψ affect performance across threshold strategies in Figure 2. Finally, we present the XGBoost learner's average threshold values across noise levels Λ to illustrate the interaction between noise levels and threshold values.

4.1. Effect of Noise Levels Λ

Figure 1 includes the average G-Mean, TPR, and TNR results for each learner and threshold combination, aggregated across all noise distributions Ψ. Beginning with the base case when there is no class noise, where $\Lambda = 0$, results show that all threshold types obtain similar G-Mean scores with respect to each learner. Unlike the G-Mean metric,

Figure 2. XGBoost Classification Performance For Noise Distributions Ψ



however, the TPR and TNR metrics show some significant differences in performance. For all three learners, $\lambda_{fmeasure}$ and λ_{gmean} obtain higher TPR rates and lower TNR rates when compared to the λ_p and λ_{np} thresholds. For example, the RF learner with $\lambda_{fmeasure}$ and λ_p thresholds obtain TPR scores of 0.8969 and 0.8869, respectively, when $\Lambda = 0$. The MLP learner sees the greatest difference, with TPR scores dropping from 0.914474 down to 0.81755 when comparing $\lambda_{fmeasure}$ to λ_p. This behavior is no surprise, as the optimal thresholds are defined by selecting thresholds that maximize performance on the training set such that $TPR > TNR$. These results for $\Lambda = 0$ are consistent with related works, and they suggest that optimizing the classification threshold on training data is an effective way to obtain desirable performance on the test set when there is no class label noise.

As noise levels Λ increase in Figure 1, G-Mean performance generally declines across each of the learners and threshold strategies. The $\lambda_{fmeasure}$ and λ_{gmean} thresholds follow a similar downward trend as noise levels increase, and generally perform the worst according to the G-Mean metric for $\Lambda \geq 0.1$. For XGBoost and MLP learners, λ_p outperforms $\lambda_{fmeasure}$ and λ_{gmean} when $\Lambda \leq 0.4$, but for the RF learner λ_p performs the worst for $\Lambda \geq 0.3$. The λ_{np} threshold consistently performs the best across all learners and noise levels according to the G-Mean metric. For the highest noise levels ($\Lambda = 0.5$), λ_{np} maintains G-Mean scores between 0.8614 and 0.9048. These G-Mean scores are significantly greater than the optimized thresholds, which have G-Mean scores in the range 0.2569–0.8361 when $\Lambda = 0.5$. Therefore, we can conclude that λ_{np} performs best across noise levels according to the G-Mean metric.

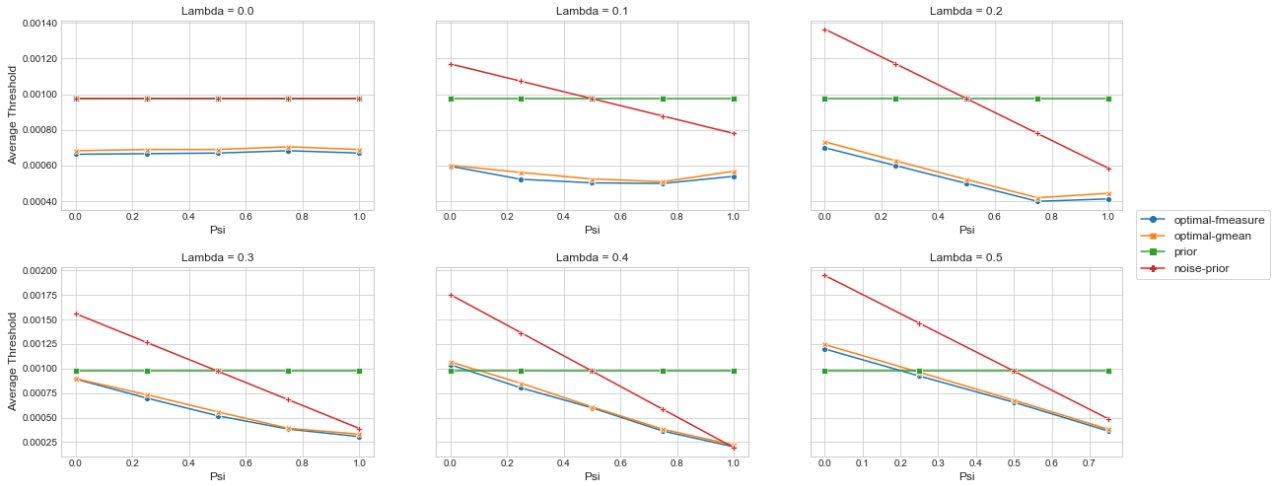
While λ_{np} clearly performs best across noise levels according to the G-Mean metric, Figure 1 illustrates significant trade-offs between TPR and TNR scores for each of the thresholds. The $\lambda_{fmeasure}$ and λ_{gmean} thresholds have increasing TPR scores and decreasing TNR scores as noise levels increase. For example, the RF learner's TPR performance increases from 0.896854 to 0.964634 as Λ increases from 0.0 to 0.5. At first glance, this appears promising, as we would generally want to maximize the TPR when detecting fraudulent providers. Unfortunately, the RF learner's TNR performance decreases from 0.861669 to 0.565188 as Λ increases from 0.0 to 0.5. Considering the size of this data set, and its highly imbalanced classes, this low TNR is unacceptable because it equates to millions of non-fraudulent providers being classified as fraudulent. This pattern of increasing TPR and decreasing TNR exists across all learners for $\lambda_{fmeasure}$ and λ_{gmean}. These thresholds, which are optimized on the noisy training sets, fail to generalize to the clean test sets. The λ_p threshold is less affected by the increasing noise levels, but it too suffers from this unbalanced class-wise performance. The λ_{np} threshold, on the other hand, maintains relatively consistent TPR and TNR performance across all noise levels and learners. As such, we would recommend using λ_{np} to obtain satisfactory and balanced performance when training data contains class noise.

4.2. Effect of Noise Distributions Ψ

Figure 2 illustrates the G-Mean, TPR, and TNR across noise distributions Ψ, averaged across all noise levels Λ using the XGBoost learner. The $\lambda_{fmeasure}$ and λ_{gmean} thresholds follow similar patterns for G-Mean, TPR, and TNR results. As Ψ increases, and the $P \rightarrow N$ noise increases, results show that G-Mean performance increases, TPR performance decreases, and TNR performance increases. The low G-Mean performance for $\Psi \leq 0.75$ is attributed to significantly lower TNR scores when noise is predominately $N \rightarrow P$. The negative samples that are mislabeled as positive samples encourage the learner to classify negative samples from the test set as positive samples, reducing the TNR. It is only when $\Psi = 1.0$, i.e. all $P \rightarrow N$ noise, that $\lambda_{fmeasure}$ and λ_{gmean} yield satisfactory performance with a high G-Mean score and an approximately balanced TPR and TNR. We suspect that the $P \rightarrow N$ noise has less effect on classification performance because there are millions of negative samples that outweigh the signal from the relatively small subset of $P \rightarrow N$ noise. These results suggest that $\lambda_{fmeasure}$ and λ_{gmean} may be acceptable when class label noise consists primarily of minority class samples being mislabeled as majority class samples. Unfortunately, the distribution of class label noise is often unknown.

Across all Ψ and performance metrics, the noise-prior threshold λ_{np} obtains stable and balanced classification performance. For any noise distribution where $\lambda_{fmeasure}$, λ_{gmean}, or λ_p outperform λ_{np}, it comes at the cost of unstable TPR or TNR performance. For example, when $\Psi = 0.0$, $\lambda_{fmeasure}$ obtains an average TPR of 0.94 and λ_{np} has an average TPR of 0.90. The TPR of the $\lambda_{fmeasure}$

Figure 3. XGBoost Average Threshold Values



is 0.04 greater than that of λ_{np}, but the TNR obtained by $\lambda_{fmeasure}$ for $\Psi = 0.0$ is only 0.79. Therefore, the TNR of $\lambda_{fmeasure}$ is very poor, especially when compared to the λ_{np} that obtained a TNR of 0.95. This trade-off of TPR for TNR exists across $\lambda_{fmeasure}$, λ_{gmean}, and λ_p. The λ_{np} threshold is the only threshold that maintains a stable and balanced classification performance across all noise distributions Ψ. Based on these results, and those from Section 4.1, we conclude that the noise-prior threshold performs best on highly imbalanced data exhibiting class label noise.

4.3. Average Threshold Values

Figure 3 illustrates the average threshold values obtained across all values of (Λ, Ψ) using the XGBoost learner. When $\Lambda = 0.0$, and there is no class label noise, $\lambda_p = \lambda_{np} = 0.00098$. The λ_p threshold has the same value across all noise levels, because it is independent of Λ and Ψ. The λ_{gmean} and $\lambda_{fmeasure}$ thresholds are approximately equal and follow similar downward trends across Ψ. The λ_{np} threshold also has a downward trend as Ψ increases, but it is consistently greater than both $\lambda_{fmeasure}$ and λ_{gmean}. The downward trend is attributed to the $P \rightarrow N$ noise associated with increasing Ψ, which in turn reduces the relative size of the positive class. In summary, the class label noise in $\mathcal{D}^n$ influences the optimal threshold selection process and decreases the values of $\lambda_{fmeasure}$ and λ_{gmean}, especially when $\Psi \leq 0.5$. These decreased values of λ explain the increased TPR and decreased TNR scores observed in Figure 2.

5. Conclusion

Related works have shown that output thresholding is an effective, and often necessary, technique for improving the classification of imbalanced data. These works recommend setting the threshold λ equal to the prior probability of the positive class, or optimizing λ on training data. To the best of our knowledge, class label noise has not been taken into consideration when evaluating these output thresholding strategies. When class label noise exists within a training data set, corruptions to class labels will change the prior probability of the positive class and influence threshold optimization processes. We address this gap in existing research, and explore robust thresholding strategies for improving the classification of highly imbalanced data that also exhibits class noise. A range of noise levels Λ and noise distributions Ψ were used to simulate the many types of class label noise found in real-world applications.

Results show that increasing levels of noise degrade the classification performance of the RF, MLP, and XGBoost learners. The noise distribution is shown to have a significant effect on classification performance, where high $P \rightarrow N$ noise yields the worst performance. Like related works, we find that the prior threshold (λ_p) and optimized thresholds ($\lambda_{fmeasure}, \lambda_{gmean}$) are both effective when there is no class noise. Across increasing levels of class noise, however, the thresholds optimized on the noisy training data yield sub-optimal G-Mean performance and unbalanced class-wise performance, with large trade-offs between TPR and TNR scores. More specifically, the $\lambda_{fmeasure}$, λ_{gmean}, and λ_p thresholds all perform very poorly in terms of the TNR for $\Lambda \geq 0.1$. Ultimately, this study found that a threshold equal to the prior probability of the positive class λ_{np} from the noisy distribution $\mathcal{D}^n$ consistently performs the best across learners, noise levels, and noise distributions. As this is the first study to consider output thresholding for imbalanced and noisy data, future works will repeat these experiments across a greater variety of data sets and classification algorithms.

References

[1] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, “Effective detection of sophisticated online banking fraud on extremely imbalanced data,”

World Wide Web, vol. 16, no. 4, pp. 449–475, Jul 2013. [Online]. Available: https://doi.org/10.1007/s11280-012-0178-0

[2] A. N. Richter and T. M. Khoshgoftaar, "Sample size determination for biomedical big data with limited labels," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, pp. 1–13, 2020.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Medical provider embeddings for healthcare fraud detection," *SN Computer Science*, vol. 2, no. 4, p. 276, 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00656-y

[4] J. M. Johnson and T. M. Khoshgoftaar, "Hcpcs2vec: Healthcare procedure embeddings for medicare fraud prediction," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, 2020, pp. 145–152.

[5] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 935–942.

[6] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119–1130, Aug 2012.

[7] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249 – 259, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608018302107

[8] J. M. Johnson and T. M. Khoshgoftaar, "Thresholding strategies for deep learning with highly imbalanced big data," *Deep Learning Applications, Volume 2*, pp. 199–227, 2021. [Online]. Available: https://doi.org/10.1007/978-981-15-6759-9_9

[9] X. Zhang, H. Gweon, and S. Provost, "Threshold moving approaches for addressing the class imbalance problem and their application to multi-label classification," in *2020 4th International Conference on Advances in Image Processing*, ser. ICAIP 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 72–77. [Online]. Available: https://doi.org/10.1145/3441250.3441274

[10] J. Van Hulse and T. M. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data & Knowledge Engineering*, vol. 68, no. 12, pp. 1513–1542, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X09001141

[11] D. Nettleton, A. Orriols-Puig, and A. Fornells Herrera, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artif. Intell. Rev.*, vol. 33, pp. 275–306, 04 2010.

[12] Centers For Medicare & Medicaid Services. (2019) Medicare provider utilization and payment data. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data

[13] J. T. Hancock and T. M. Khoshgoftaar, "Gradient boosted decision tree algorithms for medicare fraud detection," *SN Computer Science*, vol. 2, no. 4, p. 268, 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00655-z

[14] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Research*, vol. 5, pp. 2–8, 2016, big data analytics and applications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214579615000611

[15] C. L. Calvert and T. M. Khoshgoftaar, "Threshold based optimization of performance metrics with severely imbalanced big security data," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019, pp. 1328–1334.

[16] J. M. Johnson and T. M. Khoshgoftaar, "Output thresholding for ensemble learners and imbalanced big data," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021.

[17] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*, 4th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016.

[18] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.

[19] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst. Man Cybern.*, vol. 2, pp. 408–421, 1972.

[20] I. Tomek, "An experiment with the edited nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 6, pp. 448–452, 1976.

[21] S. Verbaeten and A. Van Assche, "Ensemble methods for noise elimination in classification problems," in *Proceedings of the 4th International Conference on Multiple Classifier Systems*, ser. MCS'03. Berlin, Heidelberg: Springer-Verlag, 2003, p. 317–325.

[22] T. M. Khoshgoftaar and P. Rebours, "Improving software quality prediction by noise filtering techniques," *J. Comput. Sci. Technol.*, vol. 22, pp. 387–396, 05 2007.

[23] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, "Inffc: An iterative class noise filter based on the fusion of classifiers with noise sensitivity control," *Information Fusion*, vol. 27, pp. 19 – 32, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S156625351500038X

[24] J. Sánchez, R. Barandela, A. Marqués, R. Alejo, and J. Badenas, "Analysis of new techniques to obtain quality training sets," *Pattern Recognition Letters*, vol. 24, no. 7, pp. 1015 – 1022, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865502002258

[25] I. Tomek, "Two modifications of cnn," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, 1976.

[26] J. Koplowitz and T. A. Brown, "On the relation of performance to editing in nearest neighbor rules," *Pattern Recognition*, vol. 13, no. 3, pp. 251 – 255, 1981. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320381901023

[27] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, 2019.

[28] J. Bootkrajang and A. Kabán, "Boosting in the presence of label noise," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'13. Arlington, Virginia, USA: AUAI Press, 2013, p. 82–91.

[29] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: theory, robustness to outliers, and savageboost." 01 2008, pp. 1049–1056.

[30] N. Lawrence and B. Schölkopf, "Estimating a kernel fisher discriminant in the presence of label noise." *Proceedings of the 18th International Conference on Machine Learning*, pp. 306–313, 05 2001.

[31] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 8536–8546.

[32] R. K. L. Kennedy, J. M. Johnson, and T. M. Khoshgoftaar, "The effects of class label noise on highly-imbalanced big data," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021.

[33] Office of Inspector General. (2019) Leie downloadable databases. [Online]. Available: https://oig.hhs.gov/exclusions/exclusions_list.asp

[34] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *Journal of Big Data*, vol. 6, no. 1, p. 63, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0225-0