Classifier Performance Estimation with Unbalanced, Partially Labeled Data[*]

Benjamin A. Miller BAMILLER@LL.MIT.EDU
Jeremy Vila[†] JEREMYPVILA@GMAIL.COM
Malina Kirn [‡] MALINA.KIRN@ITHRYN.COM
Joseph R. Zipkin JOSEPH.ZIPKIN@LL.MIT.EDU
MIT Lincoln Laboratory, 244 Wood St., Lexington, MA

Editors: Luís Torgo, Stan Matwin, Gary Weiss, Nuno Moniz, Paula Branco

Abstract

Class imbalance and lack of ground truth are two significant problems in modern machine learning research. These problems are especially pressing in operational contexts where the total number of data points is extremely large and the cost of obtaining labels is very high. In the face of these issues, accurate estimation of the performance of a detection or classification system is crucial to inform decisions based on the observations. This paper presents a framework for estimating performance of a binary classifier in such a context. We focus on the scenario where each set of measurements has been reduced to a score, and the operator only investigates data when the score exceeds a threshold. The operator is blind to the number of missed detections, so performance estimation targets two quantities: recall and the derivative of precision with respect to recall. Measuring with respect to error in these two metrics, simulations in this context demonstrate that labeling outliers not only outperforms random labeling, but often matches performance of an adaptive method that attempts to choose the optimal data for labeling. Application to real anomaly detection data confirms the utility of the approach, and suggests direction for future work.

Keywords: Performance estimation, class imbalance, anomaly detection, maximum likelihood, semi-supervised learning

1. Introduction

Numerous applications, in domains ranging from cyber security and biodefense to reliability engineering, involve monitoring a stream of measurements and determining when an event of interest—such as a network intrusion or hardware failure—has occurred. In these settings, operators typically work with a score computed from available observations. When the score exceeds a certain value, an in-depth investigation into the related data is performed to determine whether or not the high score indicated a true event. Such an investigation

[*] This work is sponsored by the Intelligence Advanced Research Projects Activity (IARPA) in the Office of the Director of National Intelligence (ODNI) under Air Force Contract FA8702-15-D-0001. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the United States Government.

[†] Now with Shell International Exploration & Production, Inc., 3333 Hwy 6 South, Houston, TX 77002.

[‡] Now with Ithryn, LLC.

is not performed otherwise, due to the high cost of operator and analyst effort, and the threshold is often set high to avoid alarm fatigue.

This lack of insight into the bulk of related data leads to a blind spot in the analysis of performance: no knowledge of the false negatives missed by the system. Since all scores beyond a certain value are investigated, the precision of the system is known, but the recall must be estimated based on available data and labels resulting from the deeper analysis. This is a difficult problem, especially in cases with significant class imbalance (e.g., far more scores from uninteresting data). Without the capacity to label more data, we rely on estimates in order to make informed decisions about current and future system operation.

This paper takes steps toward a framework for quantifying performance uncertainty in high-volume detection systems. Given a set of scores with labels on a subset, we quantify the ability to correctly estimate the true distributions of interesting and uninteresting data in a parameterized setting, and to discriminate between candidate distributions for background activity. From the estimated distributions, we quantify the ability to estimate both recall and the derivative of precision with respect to recall. This derivative enables informed shifting of the detection threshold in terms of the precision/recall tradeoff and the cost of additional analysis. The scenario of interest is when the only labels available are those associated with scores above the threshold, meaning we only require the labels obtained in the course of normal operation.

The remainder of this paper is organized as follows. In Section 2 we briefly review related work. Section 3 describes the problem setting in detail. Section 4 quantifies the difficulty as class balance is reduced and fewer labels are provided. Section 5 outlines the procedure we use to estimate the distribution parameters. In Section 6, we present and analyze the results of a thorough set of Monte Carlo simulations, as well as results on real anomaly detection data. Section 7 provides a summary and discusses potential next steps based on this research.

2. Related Work

This paper considers a context with three principal challenges: unbalanced classes, few labels, and uncertainty of classifier performance. Each of these factors makes the inference problem more difficult. Several techniques exist to mitigate class imbalance, including sampling, modifying the cost of false negatives versus false positives, and human-in-the-loop techniques (Weiss, 2004). These methods can be used individually, or in an ensemble to achieve better performance (Wang and Yao, 2009). Many of these methods are fully supervised. Even when all labels are present, class imbalance complicates the inference process. For example, it is hard to get a good estimate of class proportions when the class sizes are very different (Wallace and Dahabreh, 2012).

Much of the research on semi-supervised learning—where only some data are labeled—is focused on sampling the optimal set of training data. Various approaches have been proposed, such as stratified sampling (Druck and McCallum, 2011). For cases where there are both class imbalance and few labels, there has been research on optimal feature engineering (Chen et al., 2010), and techniques like active learning can be used to obtain more labels in areas of greater uncertainty (Li et al., 2012). This can be done based on one round of classification, or by analyzing the change in uncertainty when unlabeled data are given positive or negative labels, as in Juszczak and Duin (2004). Unlike in Ward et al. (2009); Elkan and Noto (2008), there are some negative labels, occurring when data from the negative class generate scores exceeding the labeling threshold.

The specific focus of this paper is local estimation of the precision–recall tradeoff curve. This is a problem when there are few labels, since there is no way to fully quantify the number of errors when labels are missing. Some methods to alleviate this issue involve cross-validating with multiple systems, such as Lamiroy and Sun's method where several classifiers are applied to the same data and the results are fused to get a probabilistic estimate of precision and recall (Lamiroy and Sun, 2013). This paper assumes a single stream of data that has already been converted into scores. We follow the analysis model of Welinder et al. (2013), but in a different context. The focus of Welinder et al. (2013) is applying a classifier trained on one dataset to a new dataset with no labels. The objective, under the assumption that labels are expensive, is to estimate the performance of the classifier on the new dataset with as few labels as possible. Our context is different: we assume that the classes are not balanced, and that there is a particular set of scores that are definitely labeled. The objective is to estimate the recall with only the given labels, and determine whether it would be beneficial to move the threshold and analyze more data.

3. Problem Model

We model the data as a set of real-valued scores, with greater scores indicating higher anticipated relevance to the analyst. There are N scores, with the ith score denoted by s_i (though the ordering can be arbitrary). We model the data as coming from a distribution mixture, where there is a background model p_0 (the distribution for uninteresting data) and a foreground model p_1 (the distribution of data for which the analyst is looking). Distributions p_0 and p_1 have parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, respectively. Some of the data have labels. We denote the set of indices for the labeled data by L, and for the unlabeled data by U. The labeled data consists of positive labels, L_1, and negative labels, L_0, such that $L_1 \cap L_0 = \emptyset$ and $L_1 \cup L_0 = L$. The distribution from which score i was drawn is denoted by y_i, which is equal to 0 if s_i was drawn from the background distribution and is 1 otherwise. Thus, for labeled data, we have $s_i \sim p_{y_i}\left(\boldsymbol{\theta}_{y_i}\right)$, whereas for unlabeled data we have $s_i \sim (1-\beta)\, p_0\left(\boldsymbol{\theta}_0\right) + \beta p_1\left(\boldsymbol{\theta}_1\right)$, where $\beta \in (0, 0.5)$ is the balance parameter.

Following Welinder et al. (2013), we optimize the coefficient estimates with respect to the joint likelihood function

$$p(\boldsymbol{s}, \boldsymbol{y}_L; \boldsymbol{\theta}, \beta) = \prod_{i \in U} \left((1-\beta) p_0(s_i; \boldsymbol{\theta}_0) + \beta p_1(s_i; \boldsymbol{\theta}_1)\right) \cdot \prod_{i \in L} \beta^{y_i} (1-\beta)^{1-y_i} p_{y_i}(s_i; \boldsymbol{\theta}_{y_i}), \tag{1}$$

where $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_1^T\right]^T$. Given the scores and labels, we fit the data to a set of candidate mixtures, as we discuss in Section 5. The resulting distributions are evaluated for their goodness of fit to the data by using a Kolmogorov–Smirnov (KS) test. The mixture determined to have the best fit is used to estimate the precision–recall tradeoff. The precision and recall are given by

$$P = \left(\beta \int_t^\infty p_1\left(s; \boldsymbol{\theta}_1\right) ds\right) \Big/ \left(\int_t^\infty p\left(s; \boldsymbol{\theta}, \beta\right) ds\right) \quad \text{and} \quad R = \int_t^\infty p_1\left(s; \boldsymbol{\theta}_1\right) ds, \tag{2}$$

respectively, with t being the detection threshold.

3.1. Labeling methods

The primary labeling method, since it models the operational scenario of interest, is labeling all data whose scores exceed a given threshold. While this can bias the inference algorithm, it makes it more likely that some of the data from the alternative distribution p_1 will be labeled, thus somewhat balancing the labels between the two classes. We consider two other methods for comparison. The first is the most common: selecting data uniformly at random. This method is problematic in our context, where class imbalance is significant (β close to 0) and it is cost prohibitive to obtain many labels. We also consider an adaptive method, inspired by active learning (see, e.g., Wang and Hua (2011)). Using this technique, the inference procedure is run several times. In the first iteration, the algorithm is completely unsupervised. After each iteration, a given number m of scores are selected to be labeled. Based on the inferred distribution $p(s; \hat{\beta}, \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1)$, labels are obtained for the m scores whose posterior probability of belonging to either class is closest to 50%. The inference procedure (described in Section 5) is then applied using all available labels. This process is repeated until the prescribed number of data are labeled, and the inference algorithm is run once more.

3.2. Metrics

We are interested in the context where an operator sets the threshold relatively high to maintain high precision, but is unsure of the missed detections. In this scenario, there are two questions we want the inference and analysis procedure to answer. We first want to know how many events of interest are missed at the current operating point. This requires an accurate estimate of recall from the observed labels and the overall data distribution. The other desirable capability is to be able to accurately determine how much of a decrease in precision would come with an increase in recall if the threshold were lowered. This requires an estimate of the derivative of precision with respect to recall, dP/dR. An accurate estimate of both of these values will give the operator the necessary information to determine how much interesting information is being ignored or discarded, and whether it is worth the increased cost in manual effort due to reduced precision to allow more data to be inspected and get a higher level of recall.

4. Estimation Impact

The two issues addressed in this paper—class imbalance and lack of labels—both make estimation of the underlying distribution more difficult, complicating the assessment of classifier performance. One way to quantify this difficulty is by considering the change in Fisher information (FI) as the data become more imbalanced or have fewer labels. The FI of a parameter θ is given by $I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\ln p\right]$. The minimum variance of an unbiased estimator for θ—independent of any other parameters[1]—is $I^{-1}(\theta)$ (Kay, 1993). For the

[1] For brevity, we ignore cross terms in this paper. As β gets small, we see much more mutual information between β and θ_0 and less between other terms.

likelihood function (1), and letting $\ell = \ln p$, we have the following second partial derivatives:

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\sum_{i \in U} \frac{1}{p^2(s_i)} \left(p_1(s_i;\theta_1) - p_0(s_i;\theta_0)\right)^2 - \frac{|L_1|}{\beta^2} - \frac{|L_0|}{(1-\beta)^2} \tag{3}$$

$$\frac{\partial^2 \ell}{\partial \theta_0^2} = \sum_{i \in U} \frac{-(1-\beta)^2}{p^2(s_i)} \left(\frac{\partial p_0}{\partial \theta_0}\right)^2 + \frac{(1-\beta)}{p(s_i)} \frac{\partial^2 p_0}{\partial \theta_0^2} + \sum_{j \in L_0} \frac{-1}{p_0^2(s_j)} \left(\frac{\partial p_0}{\partial \theta_0}\right)^2 + \frac{1}{p_0(s_j)} \frac{\partial^2 p_0}{\partial \theta_0} \tag{4}$$

$$\frac{\partial^2 \ell}{\partial \theta_1^2} = \sum_{i \in U} \frac{-\beta^2}{p^2(s_i)} \left(\frac{\partial p_1}{\partial \theta_1}\right)^2 + \frac{\beta}{p(s_i)} \frac{\partial^2 p_1}{\partial \theta_1^2} + \sum_{j \in L_1} \frac{-1}{p_1^2(s_j)} \left(\frac{\partial p_1}{\partial \theta_1}\right)^2 + \frac{1}{p_1(s_j)} \frac{\partial^2 p_1}{\partial \theta_1^2} \tag{5}$$

There are a few things to note about the equations above. First, if the distributions p_0 and p_1 are well separated (i.e., $p_0(s) \gg p_1(s)$ or $p_0(s) \ll p_1(s)$ over most of the domain), (3) acts as a weighted count of positive and negative data: weighting $1/(1-\beta)^2$ for negative labels, $1/\beta^2$ for positive, and approximately either $1/\beta^2$ or $1/(1-\beta)^2$ depending on whether the unlabeled score is more likely to be in the positive or negative distribution, respectively. This has an interpretation consistent with intuition: the information obtained by the measurements increases the most as more positive (as opposed to negative) labels are obtained. The increase in information with one new positive measurement is equivalent to $1/\beta^2$ new unlabeled or negative data, which for small β can be several orders of magnitude.

When estimating parameters of either distribution, we see another intuitive phenomenon. If β is small enough that $p \approx p_0$, then the quantity in (4) increases by the same amount whether the new observations are unlabeled or have negative labels. On the other hand, (5) increases much more when a positively labeled observation is obtained.

A numerical example is shown in Figure 1. This figure illustrates the FI for a mixture of two normal distributions ($N = 10^6$, $\mu_0 = 2$, $\mu_1 = 3.88$, and $\sigma_0^2 = \sigma_1^2 = 1$), and compares the information when labels are provided uniformly at random (i.e., in proportion to the balance of the two distributions) to when the same number of labels are provided, but with the same number of labeled data from the positive and negative class.[2] As balance decreases (the left-hand plot, holding the labeling rate constant at $|L|/N = 0.001$), the FI for parameters of the negative distribution slightly increase, while $I(\mu_1)$ and $I(\sigma_1^2)$ decrease dramatically. As β decreases (and the range of probable values narrows), $I(\beta)$ increases. When half of the labels are from the positive class, we see much slower decrease in FI for μ_0 and σ_0^2. At small values of β, the FI has a roughly linear dependence on β, as opposed to a near-quadratic dependence for randomly selected labels. The FI for β increases at an even faster rate as β decreases, while the impact on estimating μ_0 and σ_0^2 is negligible.

As the labeling rate decreases (holding $\beta = 0.001$), we see another phenomenon. For randomly labeled data, we see no change whatsoever to $I(\mu_0)$ and $I(\sigma_0^2)$. Meanwhile, $I(\beta)$, $I(\mu_1)$, and $I(\sigma_1^2)$ are all reduced by about 2 orders of magnitude, saturating at a labeling rate just above 0.001. Labeling equal numbers from both classes delays this phenomenon, enabling the same level of information to be obtained with far fewer labels. (This makes intuitive sense, since for the large labeling rates, all positive data will be labeled.)

It is also noteworthy that, whether data are labeled uniformly at random or beyond a threshold, the likelihood function remains the same, as expressed in (1) (Seaman et al.,

[2] In cases where there are fewer than $|L|/2$ positive data, all positive data are labeled and $|L| - |L_1|$ negative data are labeled.

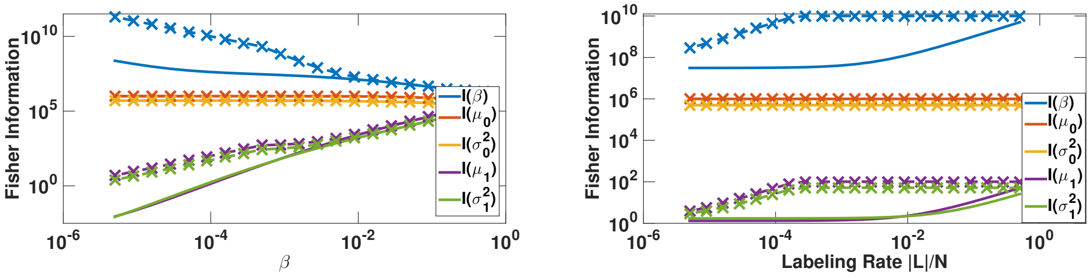

Figure 1: Fisher information of each parameter in a mixture of two normal distributions. FI is plotted with respect to the balance parameter β (left) and the labeling rate $|L|/N$ (right). Values are shown both for labeling uniformly at random (solid lines) and when half of the labeled data come from the positive class (dashed with × marks).

2013). Thus, the form of the FI is the same for either labeling method; only the parameters change. (This is not the case for the adaptive method, where data are labeled based on previous partial results. Analysis of this method is complicated and beyond the scope of this paper.)

5. Inference Algorithm

To estimate the model parameters, we use a version of expectation maximization (EM) that accounts for labels on a subset of values (Zhu and Goldberg, 2009, p. 27). This naturally interpolates between maximum likelihood (when all data are labeled) and standard EM (when there are no labels). Like traditional EM (Dempster et al., 1977), it is guaranteed to converge to a local minimum. We call this algorithm partially observable EM (POEM).

In POEM, the estimates of the parameters—$\hat{\boldsymbol{\theta}}_0$, $\hat{\boldsymbol{\theta}}_1$, and $\hat{\beta}$—are initialized to the maximum likelihood estimate based solely on the labeled data. The E-step consists of estimating the distribution of unobserved labels as $\hat{q}(\boldsymbol{y}_U) = p(\boldsymbol{y}_U|\boldsymbol{s}, \boldsymbol{y}_L; \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \hat{\beta})$, which sets the probability of a label as proportional to its likelihood under the corresponding distribution (with the most recent estimate of the parameters). The M-step maximizes the expected value of the log likelihood function with respect to $\hat{q}$. We represent the posterior probability that score i came from the positive class by

$$\gamma_i(\boldsymbol{\theta}) \triangleq \frac{1}{1 + \left(\dfrac{1-\beta}{\beta} \dfrac{p_0(s_i; \boldsymbol{\theta}_0)}{p_1(s_i; \boldsymbol{\theta}_1)}\right)} \tag{6}$$

for $i \in U$ (and either 1 or 0 for $i \in L$, depending on the label). When p_0 and p_1 are normal, this yields a simple sequence of updates:

$$N_1^{k+1} = \sum_{i=1}^{N} \gamma_i \left(\hat{\boldsymbol{\theta}}^k\right) \tag{7}$$

$$N_0^{k+1} = N - N_1^{k+1} \tag{8}$$

$$\mu_1^{k+1} = \frac{1}{N_1^{k+1}} \sum_{i=1}^{N} \gamma_i \left(\hat{\boldsymbol{\theta}}^k\right) s_i \tag{9}$$

$$\mu_0^{k+1} = \frac{1}{N_0^{k+1}} \sum_{i=1}^{N} \left(1 - \gamma_i \left(\hat{\boldsymbol{\theta}}^k\right)\right) s_i \tag{10}$$

$$(\sigma_1^2)^{k+1} = \frac{1}{N_1^{k+1}} \sum_{i=1}^{N} \gamma_i \left(\hat{\boldsymbol{\theta}}^k\right) \left(s_i - \mu_1^{k+1}\right)^2 \tag{11}$$

$$(\sigma_0^2)^{k+1} = \frac{1}{N_0^{k+1}} \sum_{i=1}^{N} \left(1 - \gamma_i \left(\hat{\boldsymbol{\theta}}^k\right)\right) \left(s_i - \mu_0^{k+1}\right)^2 \tag{12}$$

and $\beta^{k+1} = N_1^{k+1}/N$. Since, as mentioned in Section 4, the likelihood function is the same regardless of whether labels are obtained at random or based on a threshold, the algorithm does not need to be altered or reparameterized based on this condition.

Convergence results for POEM are shown in Figure 2, applied to a 2-Gaussian mixture. The parameters are $N = 10^4$, $\mu_0 = 2$ $\mu_1 = 3.88$, and $\sigma_0^2 = \sigma_1^2 = 1$. Results are shown in terms of both the mean estimate of each parameter over 100 Monte Carlo trials, and the range from the 5th to the 95th percentile of each estimate. In each case, the largest 5% of scores were labeled. In the more balanced case ($\beta = 0.3$), the initial estimates are farther from the true values than in the unbalanced case ($\beta = 0.01$), but POEM converges within a relatively small range of the actual value. In the unbalanced case, there is less variance in the estimate of the background parameters, and significantly higher variance in the estimates of the positive class distribution. There appears to be little bias in the estimates, though when β is small there is noticeable upward skew in estimates of β and σ_1^2, and downward skew in estimates of μ_1. We conclude that POEM's pessimism when it has (relatively) many positive labels is greater than its optimism when it has few.

6. Empirical Results

6.1. Monte Carlo Simulations

We ran a series of Monte Carlo simulations under a variety of conditions. In each case, the mixture consists of a background distribution that is either normal or lognormal, with mean 6.56 and variance 1, and a normal foreground distribution. We vary the labeling rate $|L|/N$ from 1/2 to 1/32 and the balance β from 0.5 to 0.005, in logarithmic steps for both variables. Each cases uses 1000 Monte Carlo trials with $N = 5000$ scores per trial.

Distribution discrimination results with $|L|/N = 1/32$ are shown in Figure 3. POEM fits the parameters of three mixtures—normal–normal, lognormal–normal, and gamma–normal—to the simulated data. When the background is normal, labeling in the tail actually

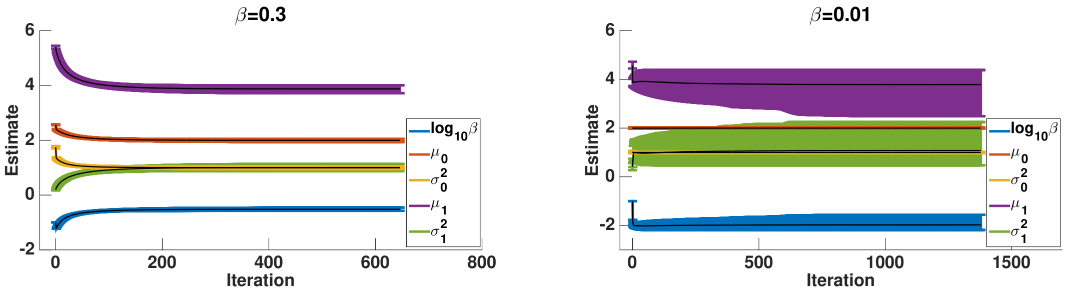

Figure 2: Convergence of POEM for a 2-Gaussian mixture over 100 trials. Shaded regions indicate the range from the 5th to 95th percentiles of estimates at a given iteration, while the black curves indicate the mean. Results are shown for balance parameters of $\beta = 0.3$ (left) and $\beta = 0.01$ (right).

hurts discrimination performance. It appears that tail labeling results in lower KS values when fitting to lognormal or gamma (but not to normal), perhaps because removing outliers yields a better estimate for the skewed distributions. The opposite is true for the lognormal background. Again, the KS statistics are similar when fit to the correct background using either method. In this case, however, labeling in the tail makes the gamma distribution a worse fit with larger KS values. This may be due to the lognormal's heavy tail: Despite the fact that both distributions have a skewed shape, more large-valued data from the background class make an exponentially tailed gamma background less likely.

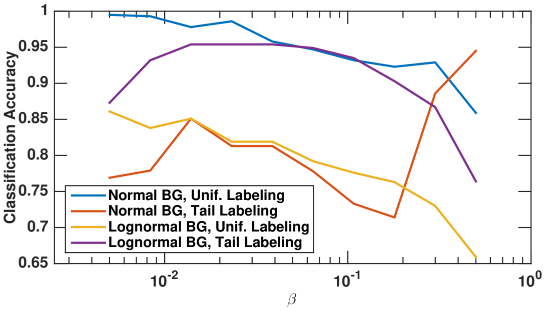

Figure 3: Accuracy when selecting a background distribution via a Kolmogorov–Smirnov test. The true background is either normal or lognormal, and is fit to normal, lognormal, and gamma distributions (with a normal foreground).

Performance estimation results are illustrated in Figure 4. The estimates of recall and dP/dR are obtained by setting a threshold in the inferred distribution mixture. The results show that—despite sometimes selecting the wrong background distribution—tail labeling yields estimates in a narrower range around the true value. The estimates are shown in terms of their 5th, 50th, and 95th percentiles, distinguishable by their vertical positions in the plots. Each plot varies the balance, the threshold where the values are estimated, or

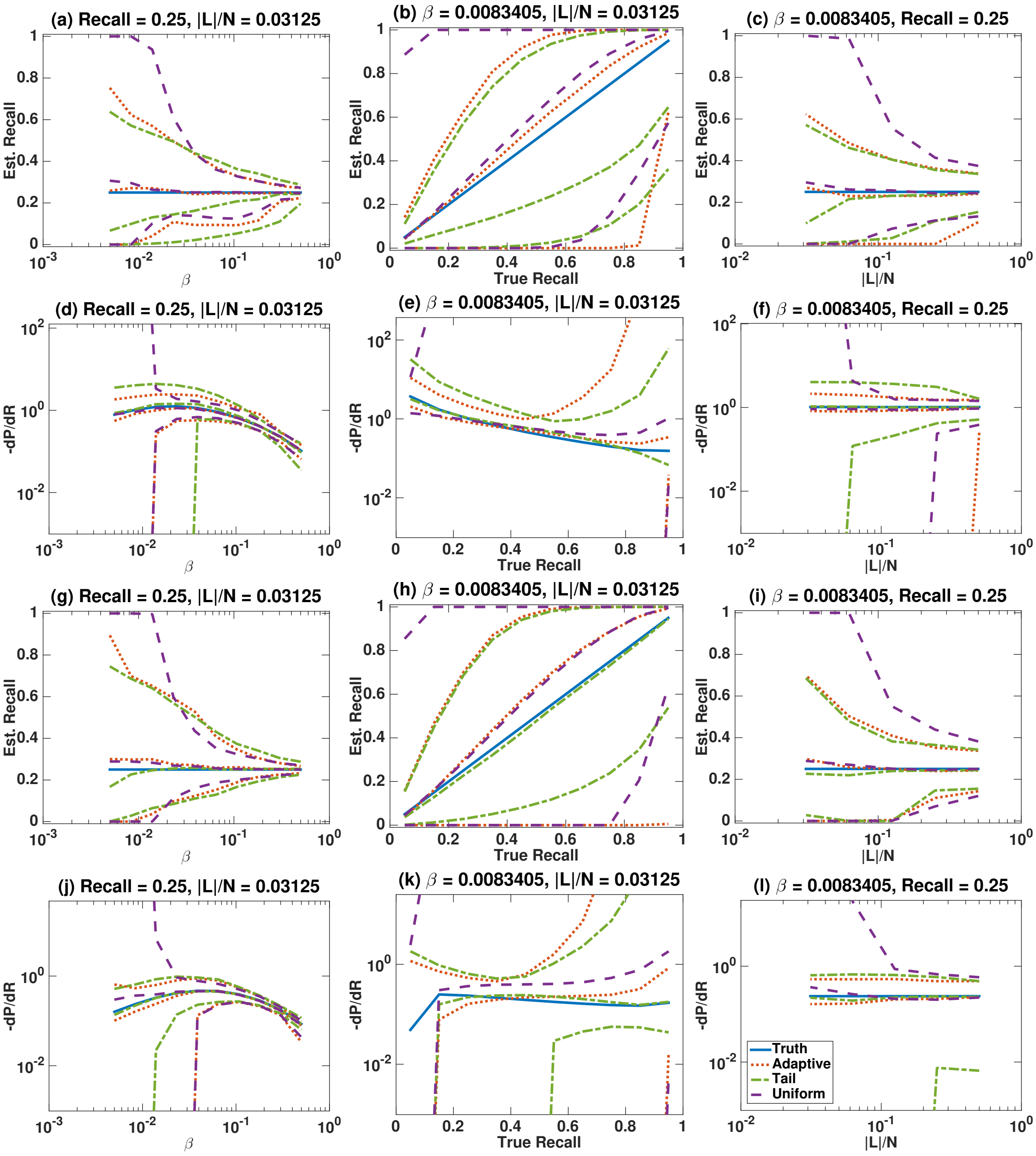

Figure 4: Monte Carlo simulation results showing estimation of recall and dP/dR. Results are shown for a normally distributed background (a–f) and a lognormal background (g–l), with respect to varying the balance parameter β (left column), the threshold defining the true recall level (center column), and the labeling rate $|L|/N$ (right column). In plots (e), (k), and (l), most of the fifth-percentile curves are not visible, as at least 5% of the data suggest a flat precision–recall curve with derivative 0, thus being at $-\infty$ on a logarithmic plot. In addition, the 95th percentile curve is not easily visible for random labeling in plots (e) and (k). This is because, in at least 5% of the trials, the slope is near-infinite.

the labeling rate. Default values are (1) $\beta \approx 0.008$, (2) the threshold where the true recall is 0.25, and (3) $|L|/N = 1/32$. For a normal background, while the median recall estimate drifts from the true value as both β and $|L|/N$ decrease, the tail labeling method has a narrower range of values than random labeling, performing similarly to the adaptive method. At relatively low (true) recall rates, the estimated recall range is also narrower with tail labeling. Tail labeling also yields more stable estimated derivatives than uniform random labeling, in particular at small values of β and $|L|/N$. While the smallest estimates for both R and dP/dR are similar across methods (all around 0), the 95th percentile estimates are much higher with random labeling. Results for a lognormal background are similar, though the estimates of dP/dR are less stable for the low values (5th percentile). This is likely due to the heavy tail of the background creating smaller values of the derivative (around -0.1), which can result in estimated distributions where dP/dR is positive.

6.2. Application Data

We applied POEM to Yahoo! Webscope's anomaly detection dataset (Yah). Two sets were selected (taken from Yahoo! service metrics), with labels of "anomaly" or "not anomaly" for each data point. These sets were selected for two reasons. First, the anomalies are sometimes change points, which is not our scenario of interest (we ignore any temporal aspects of the data). Second, many have no overlap between "anomaly" and "not anomaly" score values, leading to trivial separation. We selected the sets numbered 17 and 29, as they exhibit neither of these behaviors. For each set, we label largest 1/128 of the dataset: 12 scores out of 1424.

Performance on these two datasets is illustrated in Figure 5. In set 17, the positive class makes up about 15.9% of the scores.[3] The best fit for the data was a normal mixture, as shown in the overlaid plot in Figure 5(a). The estimated mixing parameter was $\hat{\beta} \approx 0.172$, overestimating by about a factor of about 1.077. All scores provided for training (to the right of the vertical dashed line) are positive. The precision–recall curve in Figure 5(b) shows the result of the slight overestimate of the positive class size: an underestimate of the recall level where precision starts to decrease. When the unlabeled data are given labels based on the posterior distribution, however, the estimate moves closer to the true curve. The estimate also shows dP/dR is nearly 0, and there is little cost to reducing the threshold.

Inference and estimation are more difficult in a more unbalanced dataset, as shown in Figures 5(c) and 5(d). Dataset 29's positive class has only seven measurements ($\beta \approx 0.0049$). Five of the seven positive scores fall above the threshold[4] and are marked as positive. This reduction in the balance of the classes has a substantial impact on parameter estimation. For example, a Gaussian mixture with means and variances of dataset 17, this change in β reduces the FI for μ_1 and σ_1^2 by about an order of magnitude. This dataset was also best fit by a mixture of normal distributions, but in this case $\hat{\beta} \approx 0.00365$, a much larger error. For this value of β, the expected value of the number of scores from the positive class is 5.20, just slightly more than the 5 positive scores above the threshold. This results in the highly

[3] The interested reader can also investigate set 19, which has similar class balance and similar performance.

[4] In this dataset, the labeled anomalies are anomalously *small* values, so we operate on data transformed according to $x_{\text{new}} = 2 \times 10^5 - x$.

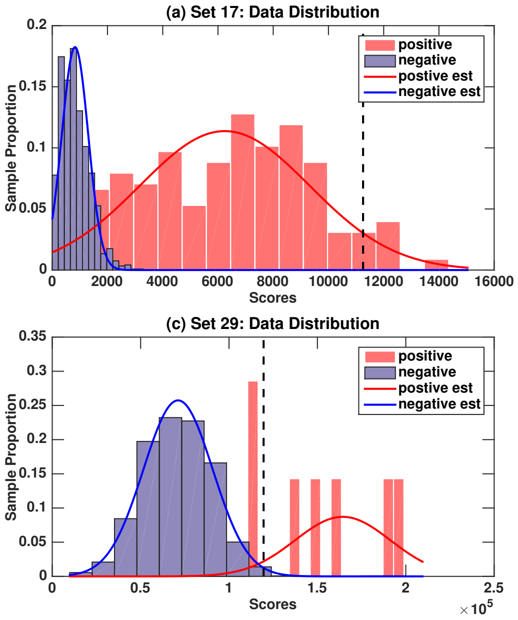

Figure 5: Results on anomaly detection data from Yahoo! Webscope. Histograms for the data in set 17 and set 29 are shown in plots (a) and (c), respectively, with the inferred distributions overlaid. The dashed line indicates the data labeling threshold. The resulting precision and recall curves are shown in plots (b) and (d), including the true curve, the curve based only on the inferred distributions, and a curve based posterior label estimates. The × marks indicate the operating point.

optimistic precision–recall curve based on the inferred distributions in Figure 5(d). Even using the posterior estimate, while the shape is similar to the true curve, it is stretched over a larger recall range to suggest that almost all positive data have been labeled, resulting in larger errors at higher recall levels. The derivative, however, is accurate: At the operating point, $dP/dR = -\infty$, implying that the precision must be reduced substantially before recall can improve, even incrementally. Adaptive labeling performs equally well on set 17, but worse on set 29, as it often misses more positive data (typically finding 2).

7. Conclusion

This paper aims to quantify the ability to estimate classifier performance with significant class imbalance and few, fixed labels. This mimics the use of detectors in practice, where data are only investigated in detail when they have raised concern for the operator. The Fisher information quantifies the increase in information gained through new labels in the smaller class—and the relative lack of information gained through new labels in the larger one. We recommend quantifying estimate quality using the error in both recall and the derivative of precision with respect to recall, which embody two important operational factors: how many events the detector finds, and how performance would change if the detection threshold were raised or lowered. Estimation performance is much better when the semi-supervised EM algorithm gets labels in the tail rather than random labeling, and often matches performance of an adaptive method. Application of this methodology to an anomaly detection dataset demonstrates both its efficacy in estimating the slope of the curve and pitfalls associated with poor estimation of the class balance.

This short study took first steps in developing a quantitative framework for classifier evaluation under these conditions. There are many possible paths for improvement on the results presented here. First, different distributions can be considered in the mixtures. The constituent distributions in the simulations always had the same variance, which could be varied in a more comprehensive study, and the results on real data suggest that a higher-entropy distribution (e.g., uniform) could be useful to model the positive class. If possible, extending the framework to a distribution-free setting—using models such as kernel density estimates—would make it more flexible. Also, the current methodology is indirect: find the best estimate for the distribution parameters in order to derive a subsequent estimate for recall and dP/dR. A more direct estimation technique focused on minimizing error in the quantities of interest would be desirable. In addition, this work (particularly the results illustrated in Fig. 4) considers the distribution of estimated metric values at a given true value. It would be even more useful to map from estimated values to a distribution of true values (likely given some prior over the true parameters). Finally, this framework will be applied to the output of a system that converts potentially high-dimensional data into a real-valued score. Integrating the proposed framework into an adaptive feedback loop with the projection method would provide the greatest utility.

Acknowledgments

The authors would like to thank Dr. John Holodnak and Mr. David J. Weller-Fahy for helpful comments on the paper.

References

A labeled anomaly detection dataset. Yahoo! Webscope. URL https://webscope.sandbox.yahoo.com/. Dataset S5, version 1.0.

X. Chen et al. Semisupervised feature selection for unbalanced sample sets of VHR images. *IEEE Geosci. and Remote Sensing Lett.*, 7(4):781–785, Oct 2010.

A. P. Dempster et al. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

G. Druck and A. McCallum. Toward interactive training and evaluation. In *Proc. ACM Int. Conf. Inform. and Knowledge Manage.*, CIKM '11, 2011.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proc. ACM Int. Conf. Knowl. Discov. Data Min.*, pages 213–220, 2008.

P. Juszczak and R. P. W. Duin. Selective sampling based on the variation in label assignments. In *Proc. Int. Conf. Pattern Recognition*, pages 375–378, 2004.

S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.

B. Lamiroy and T. Sun. Computing precision and recall with missing or uncertain ground truth. In Y.-B. Kwon and J.-M. Ogier, editors, *Graphics Recognition: New Trends and Challenges*, LNCS 7423, pages 149–162. Springer, Berlin, Heidelberg, 2013.

S. Li et al. Active learning for imbalanced sentiment classification. In *Proc. EMNLP-CoNLL*, pages 139–148, 2012.

S. Seaman et al. What is meant by 'missing at random'? *Statistical Science*, 28(2):257–268, 2013.

B. C. Wallace and I. J. Dahabreh. Class probability estimates are unreliable for imbalanced data (and how to fix them). In *Proc. IEEE ICDM*, pages 695–704, 2012.

M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol.*, 2(2):10:1–10:21, February 2011.

S. Wang and X. Yao. Diversity analysis on imbalanced data sets by using ensemble models. In *IEEE Symp. Computational Intell. and Data Mining*, pages 324–331, March 2009.

G. Ward et al. Presence-only data and the EM algorithm. *Biometrics*, 65(2):554–563, 2009.

G. M. Weiss. Mining with rarity: A unifying framework. *SIGKDD Explor. Newsl.*, 6(1): 7–19, June 2004.

P. Welinder et al. A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In *Proc. IEEE CVPR*, pages 3262–3269, 2013.

X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning.* Morgan & Claypool, 2009.