Precision/Recall on Imbalanced Test Data

Hongwei Shang[1] **Jean-Marc Langlois**[2] **Kostas Tsioutsiouliklis**[2] **Changsung Kang**[1]
[1] Walmart Global Tech [2] Yahoo Research

Abstract

In this paper we study the problem of estimating accurately the precision and recall for binary classification when the classes are imbalanced and only a limited number of human labels are available. One common strategy is to over-sample the small positive class predicted by the classifier. Rather than random sampling where the values in a confusion matrix are observations coming from a multinomial distribution, we over-sample the minority positive class predicted by the classifier, resulting in two independent binomial distributions. But how much should we over-sample? And what confidence/credible intervals can we deduce based on our over-sampling? We provide formulas for (1) the confidence intervals of the adjusted precision/recall after over-sampling; (2) Bayesian credible intervals of adjusted precision/recall. For precision, the higher the over-sampling rate, the narrower the confidence/credible interval. For recall, there exists an optimal over-sampling ratio, which minimizes the width of the confidence/credible interval. Also, we present experiments on synthetic data and real data to demonstrate the capability of our method to construct accurate intervals. Finally, we demonstrate how we can apply our techniques to Yahoo Mail's quality monitoring system.

1 INTRODUCTION

It is very important to evaluate a machine learning model's classification performance metrics accurately. In a binary classification problem it is common for the two classes to be imbalanced, and this makes an accurate evaluation harder. The motivation for this work is the following. We have deployed a set of binary mail classifiers to production. Each model classifies incoming emails into a set of categories of interest, such as finance, travel, invoice, reservation, and so on. Several of these categories are heavily imbalanced, meaning that emails belonging to these categories account for only a small percentage (less than 5% or even 1%) of all the incoming emails. Post-deployment, we would like to monitor the quality of these classifiers regularly, e.g. once every two weeks or once per month. In order to do so, we need to sample emails periodically and compute metrics like precision and recall on the sampled data. But the sampled data is expensive to generate, because it requires manually labeling by human editors. So we want to sample as few emails as possible to meet the required confidence intervals for both precision and recall. According to business needs, the estimates for precision and recall need to be accurate, i.e. within a 5% margin of error for both. Here a 5% margin of error makes the width of the confidence interval (CI) 10%.

The simplest monitoring test set would sample the data from the whole population randomly and compute the precision/recall(P/R) from the confusion matrix. Table 1 shows an example of an existing monitoring test for a binary classifier Φ based on random sampling. 5000 emails were randomly sampled from the entire email population. A confusion matrix was generated after obtaining the true labels from human editors and the predictions from the classifier (Table 1). Out of the 5000 emails, 4732 are true negatives, which shows that this class is indeed imbalanced. For this random sample, the margin of error of the estimated P/R is 5.3%/6.2%, which does not satisfy our 5% margin of error constraints.

Since this monitoring needs to be done regularly and is labor-intensive, we need to minimize the number of sampled emails. The problem is particularly challenging for those cases where the positive class is very rare. In this work we will answer the following questions: (1) How big does our sample need to be and should we over-sample the minority class? (2) Can we obtain reliable CIs for the evaluation metrics while minimizing the number of emails we need to sample and label manually?

There exists a lot of research that addresses the challenge of imbalanced classes (Fernández et al., 2017; Yuan et al.,

Proceedings of the 26$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

Table 1: An Mocked Example of Production Monitoring Test Data based on Random Sampling.

	Predicted+	Predicted−
Actual+	138	108
Actual−	22	**4732**
Precision: 86.3% ± 5.3%		
Recall: 56.1% ± 6.2%		

2018; Chawla et al., 2002; Mathew et al., 2017). Our first question is: what metrics should we use when evaluating the performance of a binary classifier? In the case of highly-imbalanced sets, some common metrics are not useful. Accuracy is not informative because there is a strong preference to always predict the negative class. Receiver Operating Characteristics (ROC) is also misleading (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006). When dealing with highly skewed data, the Precision-Recall (PR) curve is the most informative plot, compared to ROC and other metric plots (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006). Luque et al. (2019) defined several indicators to measure the impact of imbalance based on the binary confusion matrix, and showed that the Matthews Correlation Coefficient (MCC) is a good choice to demonstrate any biases in the dataset.

We now know that P/R are appropriate metrics for imbalanced data, but we still want to know how confident we are about their values. There is some prior research on assessing precision, recall, and F-score based on statistics distributions. Goutte and Gaussier (2005) assessed the confidence for P/R and F-score under random sampling strategies. Caelen (2017) further extended the work to assess the confidence for any performance indicator extracted from a confusion matrix. Both of them use a random sampling assumption, corresponding to a multinomial distribution for the confusion matrix.

Additionally, different efforts have been made to save the labeling cost in the test data. Some research works enforced parametric distributions for estimating precision-recall curves, by modeling the scores of classifiers with mixtures of densities. They focused on cases where only a small number of data is labeled (Welinder et al., 2013) or where the classes are not balanced and there is a particular set of scores labeled (Miller et al., 2018). Some other approaches use stratified sampling technique and its variations (Bennett and Carvalho, 2010; Li et al., 2019; Chen et al., 2020; Guerriero et al., 2021). The stratified sampling technique and other sampling techniques were also employed to address various evaluation problems, e.g., in information retrieval (Yilmaz et al., 2008), in evaluating generative models (Sabharwal and Xue, 2018) which adaptively estimate the whole ROC curve for a threshold class, and in evaluating multiple binary classifiers (Tripathi et al., 2020). Closer to our work, Bennett and Carvalho (2010) approximated confidence intervals based on stratified sampling, but for precision metrics only. Our work employs the stratified sampling technique by over-sampling the minority class, to take care of the skewness issue of the imbalanced data. Different from the previous work, we compute the P/R confidence intervals (CI) both in analytic forms and via simulation methods and provide an optimal over-sampling ratio to optimize the confidence of recall.

This paper presents a practical sampling strategy and a thorough analysis of the confidence of P/R metrics of a binary classifier Φ, for imbalanced data. The classifier Φ is fixed, and we focus on how to sample test data given the limited test data size, and how to deduce the approximate distributions of P/R. To the best of our knowledge, this is the first paper that thoroughly assesses the confidence of both precision and recall when applying an over-sampling strategy. (1) We illustrate the problems with inaccurate P/R metrics on randomly sampled test data. We propose to remedy this problem by over-sampling the small class, which assumes that the values in the confusion matrix come from two independent binomial distributions. (2) We formalize this problem, and propose two ways to infer approximate distributions of P/R derived from the confusion matrix. We calculate approximate confidence/credible intervals for P/R from both frequentist and Bayesian perspectives, both analytically and via simulation. Bayesian techniques allow us to take into account the intrinsic variability of the unknown parameters. (3) We recommend a sampling ratio to optimize (trade-off) the confidence of precision and recall.

This paper is organized as follows. We first formulate the problem and the over-sampling strategy in Section 2, which corresponds to two independent binomial distributions in the confusion matrix. This is a departure from the traditional approach of a multinomial distribution with random sampling. In Section 3, we first construct CIs for P/R based on a normal approximation of the binomial distribution. Next, using a Bayesian perspective, we derive the posterior predictive distribution of P/R and Bayesian credible intervals for future test data given an already observed confusion matrix and prior knowledge. In addition, we run simulations using the bootstrap method (corresponding to confidence intervals in Section 3.1) and the Monte-Carlo method (corresponding to credible intervals in Section 3.2). These approximations are tested on experiments in Section 4, followed by a real application example (Section 5) [1] . Finally, Section 6 discusses the strengths of this work and concludes.

2 PROBLEM FORMULATION

We consider the simple binary classification setting where each data point has a true label and a predicted label in $[1(+), 0(-)]$. Our goal is to estimate the P/R of the population by sampling a test set. The experimental results from

[1] This work was done when the author worked at Yahoo.

Table 2: Confusion Matrix C.

	Predicted$+$	Predicted$-$
Actual$+$	$n_{1,1}$	$n_{1,0}$
Actual$-$	$n_{0,1}$	$n_{0,0}$
	$n_{\cdot,1}$	$n_{\cdot,0}$

a sample test set are summarized in a confusion matrix C in Table 2, which reports the number of true positives ($n_{1,1}$), false positives ($n_{0,1}$), false negatives ($n_{1,0}$), and true negatives ($n_{0,0}$). We use π_{prec} and π_{recall} to denote the true Precision and Recall respectively, and $\hat{\pi}_{prec}$ and $\hat{\pi}_{recall}$ to denote the sample estimates of π_{prec} and π_{recall}.

2.1 Probabilistic Model with Random Sampling

Under random sampling assumptions, each data point in the test set is independently and identically distributed. Both precision and recall have a natural interpretation within a Bayesian probability framework (Goutte and Gaussier, 2005). In particular, precision and recall can be estimated by $\hat{\pi}_{prec} = n_{1,1}/(n_{1,1} + n_{0,1})$ and $\hat{\pi}_{recall} = n_{1,1}/(n_{1,1}+n_{1,0})$ respectively, based on the confusion matrix C (Table 2). Notice that $n_{1,1}$, $n_{1,0}$, $n_{0,1}$, $n_{0,0}$ follow a multinomial distribution.

Let $Bin(n, \pi)$ denote a Binomial distribution with n independent experiments and success probability π, and let η denote the confidence level. Note that $\eta = 5\%$ corresponds to a 95% CI. Let z_η refers to $100\eta\%$ percentile point of standard normal distribution. Given the property that marginals and conditionals of a multinomial distribution follow binomial distributions, we have:

- $n_{1,1}$ given the number of predicted positives $n_{\cdot,1}$ follows $Bin(n_{\cdot,1}, \pi_{prec})$.
- $n_{1,1}$ given the number of labeled positives $n_{1,\cdot}$ follows $Bin(n_{1,\cdot}, \pi_{recall})$.

Thus, the $100(1-\eta)\%$ approximation CIs for precision and recall are:

$$\hat{\pi}_{prec} \pm z_{1-\frac{\eta}{2}} \sqrt{(\hat{\pi}_{prec}(1-\hat{\pi}_{prec})/n_{\cdot,1}}$$

and

$$\hat{\pi}_{recall} \pm z_{1-\frac{\eta}{2}} \sqrt{\hat{\pi}_{recall}(1-\hat{\pi}_{recall})/n_{1,\cdot}}$$

respectively. When the data is very imbalanced, $n_{\cdot,1}$ is relatively much smaller than the total sample size, making the widths of the CIs very wide, especially for precision.

2.2 Probabilistic Model with Over-sampling

One intuitive solution to alleviate the class imbalance problem is to over-sample the positive data. But how much should we over-sample? First, we need to know how imbalanced the data is. The whole population is divided into $+$ve (group 1) and $-$ve (group 2) according to the prediction of the classifier. We use the ratio of the size of group 1 over the size of group 2 to denote how imbalanced the data is. This ratio will be denoted by k throughout this paper. So, for example, $k = 1/19$ means that 5% of data are predicted positive by the classifier. After k is computed, we over-sample by taking $n_{\cdot 1}$ and $n_{\cdot 0}$ random samples from groups 1 and 2 respectively. The numbers $n_{\cdot 1}$ and $n_{\cdot 0}$ are decided by an oversampling ratio s. We will discuss how to choose s in Section 3.3). Given the over-sampling ratio s, the ratio of $n_{\cdot 1}$ to $n_{\cdot 0}$ will be $k \times s$. Let v denote the total sample size (test data) for editors to judge. Then

$$n_{\cdot 1} = v \cdot k \cdot s/(k \cdot s + 1), \quad n_{\cdot 0} = v/(k \cdot s + 1) \qquad (1)$$

Unlike the random sampling scenario where both precision and recall are defined as probabilities from the multinomial distribution, over-sampling by fixing $n_{\cdot 1}$ and $n_{\cdot 0}$ divides the confusion matrix into two independent sets, the predicted positive and predicted negative sets. Let π_1 denote the ratio of true positives to predicted positives and π_0 the ratio of false negatives to predicted negatives, and let $\hat{\pi}_1$ and $\hat{\pi}_0$ be their sample estimates. Note that $\pi_1 = \pi_{prec}$ as both denote the same value. Naturally, we have

- The distribution of $n_{1,1}$ given the number of predicted positives $n_{\cdot,1}$ follows $Bin(n_{\cdot,1}, \pi_1)$.
- The distribution of $n_{1,0}$ given the number of predicted negatives $n_{\cdot,0}$ follows $Bin(n_{\cdot 0}, \pi_0)$.

Finally, the estimated precision and recall metrics from the given confusion matrix are:

$$\hat{\pi}_{prec} = n_{1,1}/n_{\cdot 1} = \hat{\pi}_1, \hat{\pi}_{Recall} = \frac{n_{1,1}}{n_{1,1}+s \cdot n_{1,0}} = \frac{1}{1+\frac{1}{k} \cdot \frac{\hat{\pi}_0}{\hat{\pi}_1}} \qquad (2)$$

3 CONFIDENCE/CREDIBLE INTERVALS BY OVER-SAMPLING

3.1 Confidence Intervals by Over-sampling

The $100(1-\eta)\%$ CIs for precision and recall can be obtained in analytic forms. Let z_η be the $100\eta\%$ percentage point of the $N(0,1)$ distribution. $\hat{\pi}_1$ is approximately normally distributed with mean π_1 and asymptotic variance $\pi_1(1-\pi_1)/n_{\cdot 1}$. Thus, the $100(1-\alpha)\%$ approximate CI for π_{prec} is:

$$\hat{\pi}_1 \pm z_{1-\frac{\eta}{2}} \sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_{\cdot 1}} \qquad (3)$$

The approximate CI of precision under over-sampling is the same as that under the random sampling scenario. But the CI of recall will be constructed very differently, because it does not follow naturally the Binomial distribution as in the random sampling scenario. Let $u = \log(\pi_0/\pi_1)$, and its estimated version $\hat{u} = \log(\hat{\pi}_0/\hat{\pi}_1)$. The estimated recall is a monotonic function of $\hat{u}$ (see Eq (2)), which is a logarithm of the ratio of two proportions from two independent binomial distributions. The variate $\hat{u}$ is approximately normally distributed with estimated mean $\log(\hat{\pi}_0/\hat{\pi}_1)$ and estimated variance $(1-\hat{\pi}_1)/(n_{\cdot 1}\hat{\pi}_1)+(1-\hat{\pi}_0)/(n_{\cdot 0}\hat{\pi}_0)$ (Katz et al., 1978). Katz et al. (1978) concluded that this method is reasonable and less conservative than two other methods proposed in the paper. Thus, the $100(1-\eta)\%$ approximate confidence interval for u is: $\log(\hat{\pi}_0/\hat{\pi}_1)\pm z_{1-\frac{\eta}{2}}\sqrt{(1-\hat{\pi}_1)/(n_{\cdot 1}\hat{\pi}_1)+(1-\hat{\pi}_0)/(n_{\cdot 0}\hat{\pi}_0)}$. Given the CI for u above, there are two ways to approximate the CI for recall. One way is to use the CI of u to transform to CI of recall. Since $\pi_{recall}=1/(1+\frac{1}{k}\exp{(u)})$ is a monotonic decreasing function of u, π_{recall}'s confidence interval will be:

$$\frac{1}{1+\frac{1}{k}\frac{\hat{\pi}_0}{\hat{\pi}_1}\exp{(\mp z_{1-\frac{\eta}{2}}\sqrt{\frac{1-\hat{\pi}_1}{n_{\cdot 1}\hat{\pi}_1}+\frac{1-\hat{\pi}_0}{n_{\cdot 0}\hat{\pi}_0}})}} \tag{4}$$

The second way is to use the delta method to first estimate the variance of recall, and then derive CI assuming normal approximation. Let $\boldsymbol{E}[\cdot]$ and $\boldsymbol{V}[\cdot]$ denote the expectation and variance of a variable. Let $f(u)=1/(1+(1/k)\exp{(u)})$ and $\hat{\pi}_{Recall}=f(\hat{u})$. The variance of the estimated recall can be approximated by

$$\boldsymbol{V}[f(\hat{u})]=\boldsymbol{V}[\hat{u}][f'(\hat{u})]^2=\boldsymbol{V}[\hat{u}]\cdot(\frac{1}{k}e^{\hat{u}})^2/(1+\frac{1}{k}e^{\hat{u}})^4$$

, where $f'(\cdot)$ denotes the derivative function of $f(\cdot)$. Thus, the $100(1-\alpha)\%$ approximate confidence interval for π_{recall} is:

$$\frac{1}{1+\frac{1}{k}\frac{\hat{\pi}_0}{\hat{\pi}_1}}\pm z_{1-\frac{\eta}{2}}\frac{\frac{1}{k}\frac{\hat{\pi}_0}{\hat{\pi}_1}}{(1+\frac{1}{k}\frac{\hat{\pi}_0}{\hat{\pi}_1})^2}\sqrt{\frac{1-\hat{\pi}_1}{n_{\cdot 1}\hat{\pi}_1}+\frac{1-\hat{\pi}_0}{n_{\cdot 0}\hat{\pi}_0}} \tag{5}$$

There is a small difference between the two CIs. For the first method, because the transformation from u to π_{recall} is not linear, the reconstituted CI will not be symmetrical around the parameter estimate, especially for two probabilities near the [0,1] boundaries. The first method (Eq (4)) is generally better, which we'll use to construct CI throughout this paper. But it is convenient to use the second one (Eq (5)) to compute margin of error, which will be used in Section 5. Note that normal approximation to the binomial distribution is a standard practice in statistics. It relies on the assumption that the four numbers in the confusion matrix are all at least 5 or 10. One alternative way for constructing CIs is using the simulation methods (Bootstrap/Monte-Carlo), yet we will not be able to compute optimal over-sampling ratio without analytical formulas.

3.1.1 Confidence Intervals by Bootstrap Approach

In the previous section we derived CI formulas for P/R. In this section we will show a nonparametric approach for deriving CIs. One well-known technique to acurately approximate the distribution of an indicator is the bootstrap method (Efron, 1992). Here, we estimate π_{prec} and π_{recall}'s distribution by bootstrapping the test data, which is equivalent to bootstrapping directly on the confusion matrix. Let Q be a large positive integer corresponding to the number of bootstrap replicas. Note that under the over-sampling setup, each replica selected from Q is obtained by random sampling $n_{\cdot,1}$ and $n_{\cdot,0}$ samples with replacement from the $n_{\cdot,1}$ +ve data points and the $n_{\cdot,0}$ −ve data points predicted by the classifier Φ. The above samplings are equivalent to doing random sampling from two Binomial distributions $Bin(n_{\cdot,1},\frac{n_{1,1}}{n_{\cdot,1}})$ and $Bin(n_{\cdot,0},\frac{n_{1,0}}{n_{\cdot,0}})$ respectively. Applying bootstrapping, these Q confusion matrices yield Q pairs of estimated P/R, which are then used to infer the distribution of P/R.

Given the population ratio k between +ve and −ve samples, and the confusion matrix of an observed test set $C(n_{1,1},n_{0,1},n_{1,0},n_{0,0})$, the bootstrap Algorithm 1 below obtains bootstrap replicas of P/R. The bootstrap algorithm BS-SAMPLER($C(n_{1,1},n_{0,1},n_{1,0},n_{0,0}),k,Q$) is described in Algorithm 1.

3.2 Bayesian Credible Intervals by Over-sampling

One main application of this work is monitoring the quality of a classifier after its initial deployment. Monitoring is especially necessary if, for example, the distribution of the content to be classified changes over time. In this monitoring scenario, we already have some estimate of the P/R based on the pre-launch test sample of the classifier. A Bayesian approach can be applied to estimate the credible intervals for P/R for a future, post-deployment monitoring test. Given the prior knowledge $\boldsymbol{\alpha}=(\alpha_{1,1},\alpha_{0,1},\alpha_{1,0},\alpha_{0,0})$ and the observed pre-launch confusion matrix $C(n_{1,1},n_{0,1},n_{1,0},n_{0,0})$ for the classifier, we like to predict the estimated P/R of future independent monitoring test samples.

In Bayesian inference, the posterior predictive distribution of future data is derived by integrating out unknown parameters π_1 and π_0. Integrating over the posterior distribution of these parameters gives a posterior predictive distribution, for future data conditional on the already-observed data C.

Throughout the rest of paper, let $\zeta_{1,1}=\alpha_{1,1}+n_{1,1}$, $\zeta_{1,0}=\alpha_{1,0}+n_{1,0}$, $\zeta_{0,1}=\alpha_{0,1}+n_{0,1}$, $\zeta_{0,0}=\alpha_{0,0}+n_{0,0}$. Then, the $100(1-\eta)\%$ approximate Bayesian credible interval for precision is:

$$\frac{\zeta_{1,1}}{\zeta_{1,1}+\zeta_{0,1}}\pm z_{1-\frac{\eta}{2}}\sqrt{\frac{\zeta_{1,1}\zeta_{0,1}(\zeta_{1,1}+\zeta_{0,1}+n_{\cdot,1})}{n_{\cdot,1}(\zeta_{1,1}+\zeta_{0,1})^2(\zeta_{1,1}+\zeta_{0,1}+1)}} \tag{6}$$

and the credible interval for recall is:

$$\frac{1}{1+\frac{1}{k}\frac{\zeta_{1,0}(\zeta_{1,1}+\zeta_{0,1})}{\zeta_{1,1}(\zeta_{1,0}+\zeta_{0,0})}\exp\mp z_{1-\frac{\eta}{2}}\sqrt{\boldsymbol{V}[\hat{u}|n_{\cdot,1},n_{\cdot,0},C,\alpha]}} \tag{7}$$

where $\boldsymbol{V}[\hat{u}|n_{\cdot,1},n_{\cdot,0},C,\alpha]$

$$=\frac{\zeta_{0,1}(\zeta_{1,1}+\zeta_{0,1}+n_{\cdot,1})}{n_{\cdot,1}\zeta_{1,1}(\zeta_{1,1}+\zeta_{0,1}+1)}+\frac{\zeta_{0,0}(\zeta_{1,0}+\zeta_{0,0}+n_{\cdot,0})}{n_{\cdot,0}\zeta_{1,0}(\zeta_{1,0}+\zeta_{0,0}+1)} \tag{8}$$

. The derivations of Eq (6) and Eq (7) are shown in Appendix. Credible intervals account for both the uncertainty in estimating parameters, plus the random variation of the

Algorithm 1 BS-SAMPLER($C(n_{1,1}, n_{0,1}, n_{1,0}, n_{0,0}), k, Q$)

```
Obtain two independent Binomial distributions $Bin(n_{\cdot,1}, \frac{n_{1,1}}{n_{\cdot,1}})$ and $Bin(n_{\cdot,0}, \frac{n_{1,0}}{n_{\cdot,0}})$.
for $q \in 1, \ldots, Q$ do
  Generate $n_{\cdot,1}$ i.i.d. samples from $Bin(n_{\cdot,1}, \frac{n_{1,1}}{n_{\cdot,1}})$, yielding $n^{(q)}_{1,1}$ positive and $n^{(q)}_{0,1}$ negative data.
  Generate $n_{\cdot,0}$ i.i.d. samples from $Bin(n_{\cdot,0}, \frac{n_{1,0}}{n_{\cdot,0}})$, yielding $n^{(q)}_{1,0}$ positive and $n^{(q)}_{0,0}$ negative data.
  Compute Prec/Recall $\hat{\pi}^{(q)}_{prec}$ and $\hat{\pi}^{(q)}_{prec}$ given the sampled confusion matrix $C^{(q)}(n^{(q)}_{1,1}, n^{(q)}_{0,1}, n^{(q)}_{1,0}, n^{(q)}_{0,0})$ above and population ratio $k$.
end
Obtain a list of Precisions $(\hat{\pi}^{(1)}_{prec}, \ldots, \hat{\pi}^{(Q)}_{prec})$ and a list of Recalls $(\hat{\pi}^{(1)}_{recall}, \ldots, \hat{\pi}^{(Q)}_{recall})$.
Construct confidence intervals of $\pi_{prec}$ and $\pi_{recall}$ from the above list of precision and recall.
```

individual values. We can see that the Bayesian credible interval has a tendency to have more variability than the confidence interval.

3.2.1 Bayesian Credible Intervals via Monte-Carlo

An alternative way to estimate the posterior predictive function is via a Monte Carlo simulation process. Both bootstrap and Monte-Carlo methods are used to obtain CIs of P/R by drawing a large number of samples from the population and computing statistics in each sample. The idea behind bootstrapping is that the sample is an estimate of the population. Monte Carlo simulation refers to the process of repeatedly creating random data from the population and computing statistics from each random sample. Caelen (2017) proposed to infer distributions of any performance indicator computed from the confusion matrix of random sampling, but without illustrating what statistics each approach corresponds to exactly.

The algorithm MC-SAMPLER($C(n_{1,1}, n_{0,1}, n_{1,0}, n_{0,0})$, k, Q) is described in Algorithm 2.

3.3 Discussion on Sample Size and Over-sampling Ratio

3.3.1 Assuming the True P/R

Although the true P/R metrics for the population data should be unknown, sometimes we are confident about the values of the true P/R metrics. Since the value of k is fixed, fixing π_{prec} and π_{recall} is equivalent to fixing π_1 and π_0.

Theorem 1. *Given the fixed test data size v and $Prec$ and $Recall$ values, the variance of the estimated recall can be minimized by choosing the over-sampling ratio $s* = \frac{1}{k}\sqrt{\frac{\Omega_0}{\Omega_1}}$, where $\Omega_1 = \pi_1/(1-\pi_1)$ and $\Omega_0 = \pi_0/(1-\pi_0)$ are the odds for probability π_1 and π_0 respectively.*

The proof of this theorem is in Appendix.

3.3.2 Assuming the Distribution of P/R

More often, practitioners are not confident about the true values of P/R for classifier Φ, but they do have some knowledge about P/R. For example, in the monitoring scenario described in Section 3.2, we have the results from the previous, pre-launch test data (confusion matrix). In that case we can reuse the Bayesian framework from Section 3.2 and we can find the optimal s^* that minimizes $\boldsymbol{V}(\hat{\pi}_{recall})$.

Theorem 2. *Given the fixed test set of size v, some prior knowledge $\boldsymbol{\alpha} = (\alpha_{1,1}, \alpha_{0,1}, \alpha_{1,0}, \alpha_{0,0})$ and the observed pre-launch confusion matrix $C(n_{1,1}, n_{0,1}, n_{1,0}, n_{0,0})$ for the classifier, the predictive variance of the estimated recall can be minimized by choosing the over-sampling ratio $s^* = \frac{1}{k}\sqrt{\Theta_0/\Theta_1}$, where $\Theta_1 = \frac{\zeta_{1,1}(\zeta_{1,1}+\zeta_{0,1}+1)}{\zeta_{0,1}(\zeta_{1,1}+\zeta_{0,1})}$ and $\Theta_0 = \frac{\zeta_{1,0}(\zeta_{1,0}+\zeta_{0,0}+1)}{\zeta_{0,0}(\zeta_{1,0}+\zeta_{0,0})}$. Note that the ζ notations were defined in Section 3.2.*

The proof of this theorem is in Appendix.

3.3.3 The Example of the Mail Classifier Φ

We introduced the example of a monitoring test for one of our binary classifiers in Section 1 and showed the confusion matrix (Table 1) on a set of 5000 random sampled emails. Now we come back to this example and would like to analyze how over-sampling will affect the P/R's CIs. To make the comparison of CIs fair, we also choose the same sample size $v = 5000$. Table 1 gives us an estimate of the imbalance ratio $k = (138+22)/(108+4732) = 0.033$ and P/R of 86.3%/56.1%.

Based on the relationship below (similarly to the estimated version Eq (2))

$$\pi_1 = \pi_{prec}, \quad \pi_0 = k \cdot \pi_1 \cdot (1/\pi_{recall} - 1) \qquad (9)$$

, we get $\pi_1 = 0.863$, $\pi_0 = 0.0223$. Given π_1 and π_0, we obtain the optimal sampling ratio for recall $s^* = 1.823$ from Theorem 1. From a Bayesian perspective, we can assume from our observations of the Confusion matrix Table 1 that the precision has a mean of 86.3% and the recall 56.1% with some variance. There is a one-to-one mapping between P/R's mean and variance and π_1 and π_0's mean and variance. Also, the mean and variance π_1 and π_0 will determine the posterior distributions $Beta(\zeta_{1,1}, \zeta_{0,1})$ and $Beta(\zeta_{1,0}, \zeta_{0,0})$, which will be set at $Beta(86.3w, 13.7w)$ and $Beta(67.5w, 2962.8w)$ to match the P/R mean. Here, we set w to different values $w =$

Algorithm 2 MC-SAMPLER($C(n_{1,1}, n_{0,1}, n_{1,0}, n_{0,0})$, k, Q)

```
Obtain distributions $Beta(n_{1,1}, n_{0,1})$ and $Beta(n_{1,0}, n_{0,0})$.
for $q \in 1, \dots, Q$ do
  Generate a random sample $\tilde{\pi}_1^{(q)}$ and $\tilde{\pi}_0^{(q)}$ from $Beta(n_{1,1}, n_{0,1})$ and $Beta(n_{1,0}, n_{0,0})$ respectively.
  Generate $n_{\cdot,1}$ i.i.d. samples from $Bin(n_{\cdot,1}, \tilde{\pi}_1^{(q)})$, yielding $n_{1,1}^{(q)}$ positive and $n_{0,1}^{(q)}$ negative data.
  Generate $n_{\cdot,0}$ i.i.d. samples from $Bin(n_{\cdot,0}, \tilde{\pi}_0^{(q)})$, yielding $n_{1,0}^{(q)}$ positive and $n_{0,0}^{(q)}$ negative data.
  Compute P/R $\hat{\pi}_{prec}^{(q)}$ and $\hat{\pi}_{recall}^{(q)}$ given the sampled confusion matrix $C^{(q)}(n_{1,1}^{(q)}, n_{0,1}^{(q)}, n_{1,0}^{(q)}, n_{0,0}^{(q)})$ above and population ratio $k$.
end
Obtain a list of Precisions $(\hat{\pi}_{prec}^{(1)}, \dots, \hat{\pi}_{prec}^{(Q)})$ and a list of Recalls $(\hat{\pi}_{recall}^{(1)}, \dots, \hat{\pi}_{recall}^{(Q)})$.
Construct credible intervals of $\pi_{prec}$ and $\pi_{recall}$ from the above list of precision and recall.
```

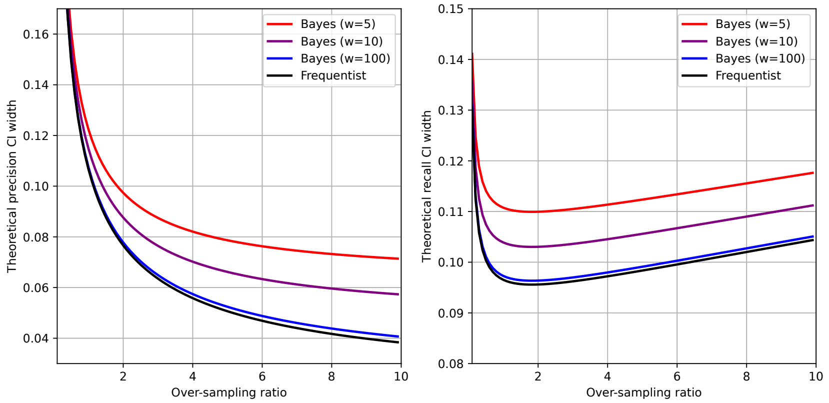

Figure 1: Plots of widths for confidence intervals (black line) and credible intervals with different w values; Left: precision; Right: recall.

$5, 10, 100$ to represent different variances; and the corresponding $s^* = 1.821, 1.822, 1.823$ based on Theorem 2. Figure 1 shows how the CI width changes for precision (left panel) and recall (right panel) as a function of the over-sampling ratio s. The larger the value of w, the smaller the standard deviation of π_1 and π_0.

For precision (Figure 1 Left), the widths for both confidence interval and credible interval are always getting smaller and smaller as the over-sampling ratio s increases. For a fixed number of test data size v, bigger s will lead to bigger $n_{\cdot,1}$ (Eq 1), thus resulting in a narrower CI width of precision (proportional to $1/\sqrt{n_{\cdot,1}}$ in Eq 3 and 6). For recall (Figure 1 Right), the widths for both confidence interval and credible interval decrease until s approaches s^*, and then they slowly increase as s continues to increase. The values of s^* are very close for both frequentist and Bayesian distributions with different w. For both precision and recall curves, the CI widths for Bayesian are always wider than those for frequentist. And, the larger w gets, the closer the Bayesian curve gets to the frequentist curve (black plots in Figure 1).

4 EXPERIMENTS

In this section, we use simulated data to assess our analytical method's ability to construct accurate confidence/credible intervals against the simulation methods (Bootstrap and Monte Carlo). The parameters of the experiment are the imbalance ratio k for classifier Φ's prediction, the true precision π_{prec} and recall π_{recall} of Φ over the population data, the sample size v, and the over-sampling ratio s. Here, k is chosen from $\{1/20, 1/100\}$; precision π_{prec} is fixed at 0.9; recall π_{recall} is chosen from $\{0.7, 0.9\}$; the over-sampling ratio is chosen from $\{1, 2, 5\}$. The sample size v is set to 5000 when $k = 1/20$, and we increase it to 10,000 when $k = 1/100$ in order to satisfy the condition of normal approximation to the binomial distribution. Given a combination of k, π_{prec} , π_{recall} , v, s, we can obtain the approximate coverage probability. For each combination of k, P/R, v, s (corresponding to each row in Table 3), we generate $A = 1000$ samples. For each generated sample, we construct

- confidence intervals $U^P_{freq,\boldsymbol{N}}$ and $U^R_{freq,\boldsymbol{N}}$ by normal approximation (Eq (3) and Eq (4));
- confidence intervals $U^P_{freq,B}$ and $U^R_{freq,B}$ by the bootstrap method (Algorithm 1);
- credible intervals $U^P_{bys,\boldsymbol{N}}$ and $U^R_{bys,\boldsymbol{N}}$ by normal approximation (Eq (6) and Eq (7));

- credible intervals $U^R_{bys,MC}$ and $U^R_{bys,MC}$ by the Monte-Carlo method (Algorithm 2).

for both precision and recall. The bootstrap size and the Monte-Carlo size Q is set to 1000, and all confidence/credible intervals are obtained at the 5% level. Finally, we use Γ to denote the approximate coverage probability for the confidence/credible intervals above.

Experimental design

- For every $a \in 1, \ldots, A$, repeat the following steps:
    - Generate a sample with sample size v and over-sampling ratio s from a population for which the classifier's P/R values are π_{prec}/π_{Recall}, form a confusion matrix C^a, and estimate P/R as $\hat{\pi}^a_{prec}$/$\hat{\pi}^a_{recall}$ from Eq (2);
    - Independently, generate another sample with the sample size v and over-sampling ratio s from the same population as above, and calculation P/R at $\tilde{\pi}^a_{prec}$/$\tilde{\pi}^a_{recall}$. We will use $\tilde{\pi}^a_{prec}$/$\tilde{\pi}^a_{recall}$ for calculating coverage probability of Bayesian credible intervals, since credible intervals are derived based on posterior predictive distribution and they are to predict P/R for future test data from the same population.
    - Given $\hat{\pi}^a_{prec}$/$\hat{\pi}^a_{recall}$, obtain both confidence intervals $U^{P,(a)}_{freq,\boldsymbol{N}}$ and $U^{R,(a)}_{freq,\boldsymbol{N}}$, and credible intervals for $U^{P,(a)}_{bys,\boldsymbol{N}}$ and $U^{R,(a)}_{bys,\boldsymbol{N}}$ of both P/R from normal approximations;
    - Given C^a, obtain confidence intervals $U^{P,(a)}_{freq,B}$ and $U^{R,(a)}_{freq,B}$ by BS-SAMPLER(C^a, k, Q);
    - Given C^a, obtain credible intervals $U^{R,(a)}_{bys,MC}$ and $U^{R,(a)}_{bys,MC}$ by MC-SAMPLER(C^a, k, Q) assuming non-informative priors.
- Approximate coverage probabilities for all the above intervals as follows:

$$\begin{aligned}
&\Gamma^P_{freq,\boldsymbol{N}} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\pi_{prec} \in U^P_{freq,\boldsymbol{N}}\},\\
&\Gamma^P_{freq,B} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\pi_{prec} \in U^P_{freq,B}\},\\
&\Gamma^P_{bys,\boldsymbol{N}} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\tilde{\pi}^a_{prec} \in U^P_{bys,\boldsymbol{N}}\},\\
&\Gamma^P_{byes,MC} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\tilde{\pi}^a_{prec} \in U^P_{bys,MC}\},\\
&\Gamma^R_{freq,\boldsymbol{N}} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\pi_{recall} \in U^R_{freq,\boldsymbol{N}}\},\\
&\Gamma^R_{freq,B} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\pi_{recall} \in U^R_{freq,B}\},\\
&\Gamma^R_{bys,\boldsymbol{N}} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\tilde{\pi}^a_{recall} \in U^R_{bys,\boldsymbol{N}}\},\\
&\Gamma^R_{bys,MC} = A^{-1}\textstyle\sum_{a=1}^{A} \mathbf{1}\{\tilde{\pi}^a_{recall} \in U^R_{bys,MC}\}
\end{aligned}$$

The experimental results are shown in Table 3. This table shows the coverage probabilities for precision/recall at various values of v, k, π_{prec}, π_{recall}, and s. Coverage probability gives the proportion of times the true precision(π_{prec})/recall(π_{recall}) was in the interval. We see that, overall, the confidence/credible intervals hold their levels reasonably well for both the analytical method and simulation methods, since most coverage probabilities are close to the expected value 95% (corresponds to $\eta = 5\%$ confidence level). Note that for cases where $v = 10000$, $k = 1/100$, and $s = 1$, slightly low coverage probabilities are observed for $\Gamma^P_{freq,\boldsymbol{N}}$. This is expected since $n_{\cdot,1} = 100$ and $n_{\cdot,1}\pi_{prec} = 10$, it lies on the borderline of satisfying the normal approximation assumption of a binomial distribution that both sample size$\times$success probability and sample size$\times(1-$success probability$)$ should be at least 10. When the normal approximation assumption is not satisfied, one can either obtain "exact" confidence intervals based on inverting the binomial test, or adjust the normal-approximated confidence intervals by adding two "successes" and two "failures" to the sample (Agresti and Coull, 1998). Since this is not the focus of our paper, we use normal approximations to compute confidence/credible intervals analytically throughout this paper.

Here, four methods (frequentist analytical, frequentist bootstrap, Bayesian analytical, Bayesian MC) were used for computing Prec/Recall. We would like to give comparisons (a) between analytical methods versus simulation methods and (b) frequentist methods versus Bayesian methods.

- **Analytical vs Simulation** The analytical derivation of confidence/credible intervals is complex and needs approximation, while using a simulation process to estimate the distribution is easy and accurate. Yet, the superiority of the analytical method is that we can have exact formulas and correspondingly derive the optimal over-sampling ratio for recall, which cannot be computed through simulation process.
- **Frequentist vs Bayesian** The Bayesian framework allows us to inject prior knowledge into the posterior. We can observe that the CIs generated by the Bayesian method are always higher than the CIs generated by the frequentist method. This can be explained by the fact that the Bayesian method takes the variability of the unknown parameters into account, whereas the frequentist method assumes them to be fixed.

5 APPLICATION

We have some highly imbalanced classifiers deployed in Yahoo Mail production. Post-deployment, we wish to monitor the quality of those models at regular intervals. Based on business needs, the estimates for precision and recall need to be accurate, so that the precision and recall are within 5% margin of error. Given a limited budget for monitoring, the question becomes how big of a sample we should get and whether we should over-sample the small class or not?

Note that before the model is deployed to production we already have evaluated P/R from a pre-launch test set. Hence we have a reasonable estimate of P/R and the ratio k of

Table 3: Coverage probabilities (%) from 1000 samples at various values of v, k, π_{prec}, π_{recall}, and s.

v	$1/k$	π_{prec}	π_{Recall}	s	$\Gamma^P_{freq,\boldsymbol{N}}$	$\Gamma^P_{freq,B}$	$\Gamma^P_{bys,\boldsymbol{N}}$	$\Gamma^P_{bys,MC}$	$\Gamma^R_{freq,\boldsymbol{N}}$	$\Gamma^R_{freq,B}$	$\Gamma^R_{bys,\boldsymbol{N}}$	$\Gamma^R_{bys,MC}$
5000	20	0.9	0.9	1	94.2	95.2	95.2	95.3	95.2	95.0	95.4	96.2
5000	20	0.9	0.7	1	94.1	95.6	95.0	95.4	95.3	94.2	93.6	93.0
5000	20	0.9	0.9	2	94.0	94.5	95.0	95.7	94.9	94.1	93.0	93.5
5000	20	0.9	0.7	2	93.6	93.3	94.1	94.8	93.6	92.9	95.2	94.7
5000	20	0.9	0.9	5	95.5	95.4	93.7	94.1	94.9	93.5	94.8	94.1
5000	20	0.9	0.7	5	94.5	94.7	95.0	95.5	94.8	94.8	95.6	95.6
10000	100	0.9	0.9	1	92.7	94.4	94.4	96.1	95.7	93.3	93.6	93.7
10000	100	0.9	0.7	1	93.5	95.7	95.3	95.8	95.1	95.0	95.5	95.6
10000	100	0.9	0.9	2	95.8	95.1	94.1	94.8	95.4	92.7	93.2	93.7
10000	100	0.9	0.7	2	94.9	94.4	93.7	94.7	96.1	95.8	94.7	94.2
10000	100	0.9	0.9	5	92.9	93.8	95.0	95.7	95.1	93.9	94.0	94.0
10000	100	0.9	0.7	5	94.0	94.8	94.1	95.4	95.0	94.6	95.5	95.4

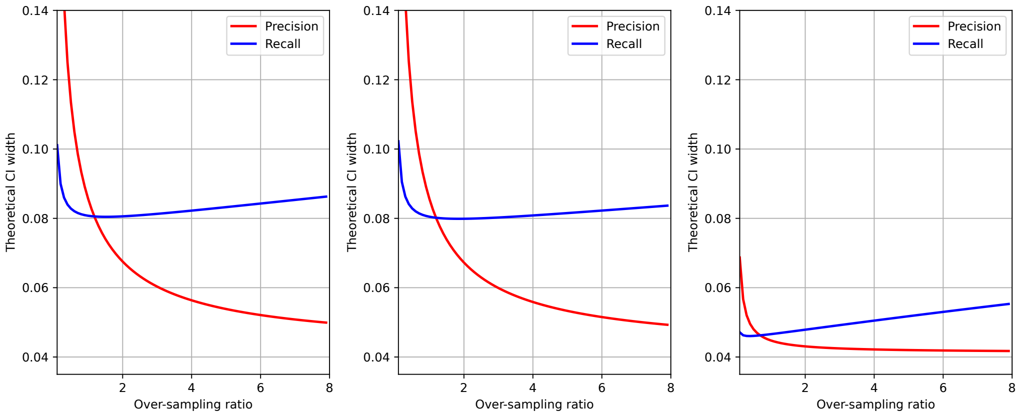

Figure 2: Plots of credible intervals' widths against the over-sampling ratios; Left: classifier Φ_1; Middle: classifier Φ_2; Right: classifier Φ_3.

positive/negative data in the population. Here, we focus on monitoring the model post-deployment, and we will estimate the minimum number of samples needed to reach a desired CI width. Table 4 shows typical examples for a few classifiers Φ_1, Φ_2, and Φ_3. For each classifier, we plot the relationship between the over-sampling ratio s and the width of the confidence intervals for both precision and recall (see Figure 2) by fixing the sample size $v = 10000$. Note that the relationship is not affected by the sample size, since the CI's width is proportional to $1/\sqrt{v}$ (see Eq 3 and Eq 5). When the data is very imbalanced (Φ_1 and Φ_2 in this application), over-sampling can greatly increase the precision confidence.

Given k, π_{prec} and π_{recall}, we first obtain π_1 and π_0 from Eq (9); then we use Theorem 1 to compute the optimal over-sampling ratio s^* for recall. If $s^* < 1$ then we set $s^* = 1$. Based on CIs for P/R in Eq (3) and Eq (5), we can obtain the margin of error for P/R (denoted by e_P and e_R respectively) as below:

$$\begin{aligned} e_P =& z_{1-\frac{\eta}{2}} \sqrt{\frac{\pi_1(1-\pi_1)}{n_{\cdot,1}}}, \\ e_R =& z_{1-\frac{\eta}{2}} \frac{\frac{1}{k}\frac{\pi_0}{\pi_1}}{(1+\frac{1}{k}\frac{\pi_0}{\pi_1})^2} \sqrt{\frac{1-\pi_1}{n_{\cdot,1}\pi_1} + \frac{1-\pi_0}{n_{\cdot,0}\hat{\pi}_0}} \end{aligned} \tag{10}$$

Note that $\hat{\pi}_1$ and $\hat{\pi}_0$ in Eq (3) and Eq (5) were replaced by π_1 and π_0 since we are using π_1 and π_0 to estimate the CIs.

Our goal is to control both $e_P <= b$ and $e_R <= b$, where the margin of error $b = 5\%$. Plugging in k, π_1, and π_0 into Eq (10), we get:

$$\begin{aligned} & n_{\cdot,1} \geq \pi_1(1-\pi_1)(\frac{z_{1-\frac{\eta}{2}}}{b})^2, \\ & \frac{1-\pi_1}{n_{\cdot,1}\pi_1} + \frac{1-\pi_0}{n_{\cdot,0}\hat{\pi}_0} \leq (\frac{b}{z_{1-\frac{\eta}{2}}})^2 \frac{(1+\frac{1}{k}\frac{\pi_0}{\pi_1})^4}{(\frac{1}{k}\frac{\pi_0}{\pi_1})^2} \end{aligned} \tag{11}$$

First we compute the minimum size $n_{\cdot,1}$ and $n_{\cdot,0}$ needed in order to control e_R. For example, take classifier Φ_1. Since $s^* = 1.51$, as shown in Table 4, we replace $n_{\cdot,0}$ with $n_{\cdot,1}/(ks^*)$ in the 2nd inequality in Eq (11), and obtain minimum $n_{\cdot,1} = 307$ and $n_{\cdot,0} = 4410$. Then, we plug

Table 4: Production Models Monitoring.

	P/R	k	π_0	s^*	$n_{\cdot,1}$	$n_{\cdot,0}$	v
Φ_1	0.79/0.67	0.046	0.0179	1.51	307	4410	4717
Φ_2	0.86/0.56	0.033	0.0223	1.85	265	4340	4605
Φ_3	0.90/0.66	0.458	0.212	1.00	141	306	447

$n_{\cdot,1} = 307$ into the precision inequality Eq (11) and it satisfies it. If the inequality is not satisfied, $n_{\cdot,1}$ and $n_{\cdot,0}$ will need to be adjusted. Similarly, we obtain the required size $n_{\cdot,1}$ and $n_{\cdot,0}$ for classifier Φ_2 and Φ_3 shown in Table 4.

6 CONCLUSIONS

We presented fundamental improvements in the accuracy of P/R metrics by proposing over-sampling of predicted positive data for imbalanced binary classification. Unlike the random sampling scenario where both precision and recall are defined as probabilities from the multinomial distribution, over-sampling naturally divides the confusion matrix into two independent sets, the predicted positive and predicted negative sets. We did a thorough analysis, including formulating the problem and calculating the statistics. We derived approximate confidence intervals analytically, and credible intervals from posterior predictive distribution by injecting knowledge from priors and previously observed data. We tested the derived formulas' capacity to construct accurate intervals, and demonstrated it on a real application example. While over-sampling is not a novel idea, to the best of our knowledge, this is the first work that studies the effects of oversampling on precision and recall in detail, and suggests specific over-sampling ratios.

Our intuition about over-sampling is that the more imbalanced the data, the more we should over-sample. By over-sampling, we obtain more data for the positive set, thus becoming more confident about the precision. For recall, the CI width reaches its minimum at an optimal over-sampling ratio s^*, and then starts increasing as s gets bigger. However, the slope of the graph increases very slowly – the more imbalanced the data, the flatter the slope. Thus, generally, it is a good idea to over-sample positive data since the precision can benefit a lot, while hurting recall very little. But if we want to be precise, we can use the imbalance ratio k and an estimate of P/R. Then we can calculate the trade-off between precision and recall, using the formulas we provided. Note that different applications may prioritize prec/recall differently. We focus on recall in our application since our production application requires both the precision and recall's CI width within 10%.

Lastly, we would like to emphasize that our work focuses on saving the labeling cost of test data rather than training data. In many applied settings, it is required to constantly monitor a previously trained classifier by evaluating the latest batch of data, due to the potential data drift which may hurt the classifier's performance. This requires periodic labeling of newly collected test data in order to detect whether the classifier can still be relied on. Our work is very useful in such settings. Some readers may be concerned about the usefulness of saving labeling effort for evaluation given the fact that maybe millions of labeled data are needed for model training anyway. Note that for model training, in many cases the majority of the training data labels are pseudo labels. For example, consider the email classifier in (Kang et al., 2022), where pseudo labels for training data were created in multiple ways. In this paper we use models that have been trained on a fairly large number of samples and that labeling effort for training data is not considered here.

Acknowledgements

We would like to thank all reviewers for their constructive suggestions to improve the paper. We are grateful to the Mail science team, engineering team, and data team at Yahoo for working closely on this project. We would like to thank the annotator team for spending hours judging emails. We would also like to thank Dr. Yong Liu and Dr. Datong Chen for the inspired discussions to make the paper better.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Bennett, P. N. and Carvalho, V. R. (2010). Online stratified sampling: evaluating classifiers at web-scale. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1581–1584.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, 81(3-4):429–450.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, J., Wu, Z., Wang, Z., You, H., Zhang, L., and Yan, M. (2020). Practical accuracy estimation for efficient deep neural network testing. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 29(4):1–35.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Fernández, A., del Río, S., Chawla, N. V., and Herrera, F. (2017). An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2):105–120.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer.

Guerriero, A., Pietrantuono, R., and Russo, S. (2021). Operation is the hardest teacher: estimating dnn accuracy looking for mispredictions. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 348–358. IEEE.

Kang, C., Shang, H., and Langlois, J.-M. (2022). Classifying emails into human vs machine category. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7069–7077.

Katz, D., Baptista, J., Azen, S., and Pike, M. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, pages 469–474.

Li, Z., Ma, X., Xu, C., Cao, C., Xu, J., and Lü, J. (2019). Boosting operational dnn testing efficiency through conditioning. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 499–509.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231.

Mathew, J., Pang, C. K., Luo, M., and Leong, W. H. (2017). Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE transactions on neural networks and learning systems*, 29(9):4065–4076.

Miller, B. A., Vila, J., Kirn, M., and Zipkin, J. R. (2018). Classifier performance estimation with unbalanced, partially labeled data. In *International Workshop on Cost-Sensitive Learning*, pages 4–16.

Sabharwal, A. and Xue, Y. (2018). Adaptive stratified sampling for precision-recall estimation. In *UAI*, pages 825–834.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Smith, D. (1983). Algorithm AS 189: Maximum likelihood estimation of the parameters of the beta binomial distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 32(2):196–204.

Tripathi, R., Jagannathan, S., and Dhamodharaswamy, B. (2020). Estimating precisions for multiple binary classifiers under limited samples. In *ECML/PKDD (4)*, pages 240–256.

Welinder, P., Welling, M., and Perona, P. (2013). A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3269.

Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610.

Yuan, X., Xie, L., and Abouelenien, M. (2018). A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. *Pattern Recognition*, 77:160–172.

SUPPLEMENTARY MATERIALS

A BAYESIAN CREDIBLE INTERVALS DERIVATIONS

The simplest Bayesian inference for a binomial parameter π uses $Beta$ distribution as the prior. The probability density function of $Beta(\gamma_1, \gamma_0)$ for π is proportional to $\pi^{(\gamma_1-1)}(1-\pi)^{(\gamma_0-1)}$. The $Beta$ distribution has

$$\boldsymbol{E}[\pi] = \gamma_1/(\gamma_1+\gamma_0), \qquad \boldsymbol{V}[\pi] = \gamma_1\gamma_0/[(\gamma_1+\gamma_0)^2(\gamma_1+\gamma_0+1)]$$

The beta distribution is the conjugate prior distribution for inference about a binomial parameter. Let variable Y follow a binomial distribution $Bin(n, \pi)$, and y denote a realization of Y. With a prior distribution $Beta(\gamma_1, \gamma_0)$, the posterior distribution for binomial parameter π in $Bin(n, \pi)$ is a $Beta(y+\gamma_1, n-y+\gamma_0)$.

As defined in Section 3.2, $\boldsymbol{\alpha} = (\alpha_{1,1}, \alpha_{0,1}, \alpha_{1,0}, \alpha_{0,0})$ represents the prior knowledge, so, naturally π_1 and π_0 have $Beta$ priors $Beta(\alpha_{1,1}, \alpha_{0,1})$ and $Beta(\alpha_{1,0}, \alpha_{0,0})$ respectively. One commonly used prior sets $\alpha_{1,1} = \alpha_{0,1} = \alpha_{1,0} = \alpha_{0,0} = 0$, which corresponds to a noninformative conjugate prior for a binomially distributed random variable.

The observed pre-launch confusion matrix is $C(n_{1,1}, n_{0,1}, n_{1,0}, n_{0,0})$, thus $n_{1,1}$ and $n_{1,0}$ are the realizations of $Bin(n_{\cdot,1}, \pi_1)$ and $Bin(n_{\cdot,0}, \pi_0)$ respectively. Let $\hat{\pi}_1 = n_{1,1}/n_{\cdot 1}$ and $\hat{\pi}_0 = n_{1,0}/n_{\cdot 0}$ be estimates of π_1 and π_0 from the observed pre-launch confusion matrix.

Given the prior $\boldsymbol{\alpha}$ and the observed pre-launch confusion matrix $C(n_{1,1}, n_{0,1}, n_{1,0}, n_{0,0})$, the posterior distributions for π_1 and π_0 are $Beta(\alpha_{1,1}+n_{1,1}, \alpha_{0,1}+n_{0,1})$ and $Beta(\alpha_{1,0}+n_{1,0}, \alpha_{0,0}+n_{0,0})$ respectively. For simplicity, we denote these as $Beta(\zeta_{1,1}, \zeta_{0,1})$ and $Beta(\zeta_{1,0}, \zeta_{0,0})$.

For future monitoring test samples, given the same k (due to the same classifier Φ), suppose that we would like to sample a test set of size v and over-sampling ratio s. Then the number of +ve and −ve samples will be $n_{\cdot,1} = \frac{k \cdot s}{k \cdot s+1} v$ and $n_{\cdot,0} = \frac{1}{k \cdot s+1} v$. Thus, we can compute the posterior predictive distribution for $\hat{\pi}_1$ and $\hat{\pi}_0$ for the future monitoring test sample. Let $\boldsymbol{\alpha}_1 = (\alpha_{1,1}, \alpha_{0,1})$ and $\boldsymbol{\alpha}_0 = (\alpha_{1,0}, \alpha_{0,0})$. The posterior predictive distribution is formulated as

$$f(\hat{\pi}_g | n_{\cdot,g}, C, \boldsymbol{\alpha}_g) = \int_{\pi_g} f_g^b(\hat{\pi}_g | n_{\cdot,g}, \pi_g) f_g^\beta(\pi_g | C, \boldsymbol{\alpha}_g) d\pi_g \qquad g \in [0,1],$$

where $f_g^b(\hat{\pi}_g | n_{\cdot,g}, \pi_g)$ is the binomial distribution and $f_g^\beta(\pi_g | C, \alpha)$ is the $Beta$ posterior distribution of π_g ($g \in [0,1]$). It shows that this posterior predictive function is the expectation of the conditional probability density function $f_g^b(\hat{\pi}_g | n_{\cdot,g}, \pi_g)$ over the posterior distribution $f_g^\beta(\pi_g | C, \alpha)$.

For this compound distribution, $n_{\cdot,g}\hat{\pi}_g \sim Bin(n_{\cdot,g}, \pi_g)$, where π_g is a random variable following $Beta(\zeta_{1,g}, \zeta_{0,g})$. Then $n_{\cdot,g}\hat{\pi}_g$ follows a beta-binomial distribution (Smith, 1983).

$$\boldsymbol{V}[\hat{\pi}_g | n_{\cdot,g}, C, \alpha] = \frac{\zeta_{1,g}\zeta_{0,g}(\zeta_{1,g}+\zeta_{0,g}+n_{\cdot,g})}{n_{\cdot,g}(\zeta_{1,g}+\zeta_{0,g})^2(\zeta_{1,g}+\zeta_{0,g}+1)} \tag{12}$$

The above variance (Eq 12) could also be derived from the law of total variance as below.

Variance of posterior predictive distribution By the law of total variance,

$$\boldsymbol{V}[\hat{\pi}_g|n_{\cdot,g},C,\alpha] = \boldsymbol{E}_{\boldsymbol{\zeta}_g}[\boldsymbol{V}[\hat{\pi}_g|n_{\cdot g},\pi_g]] + \boldsymbol{V}_{\boldsymbol{\zeta}_g}[\boldsymbol{E}[\hat{\pi}_g|n_{\cdot g},\pi_g]], \quad g \in [0,1].$$

From the binomial distribution,

$$\boldsymbol{E}[\hat{\pi}_g|n_{\cdot,g},\pi_g] = \pi_g, \quad \boldsymbol{V}[\hat{\pi}_g|n_{\cdot,g},\pi_g] = \pi_g(1-\pi_g)/n_{\cdot,g}, \quad g \in [0,1]$$

Thus,

$$\boldsymbol{V}[\hat{\pi}_g|n_{\cdot,g},C,\alpha] = \boldsymbol{E}_{\boldsymbol{\zeta}_g}[\frac{\pi_g(1-\pi_g)}{n_{\cdot,g}}|n_{\cdot,g}] + \boldsymbol{V}_{\boldsymbol{\zeta}_g}[n_{\cdot,g}\pi_g|n_{\cdot,g}] = \frac{\zeta_{1,g}\zeta_{0,g}(\zeta_{1,g}+\zeta_{0,g}+n_{\cdot,g})}{n_{\cdot,g}(\zeta_{1,g}+\zeta_{0,g})^2(\zeta_{1,g}+\zeta_{0,g}+1)}$$

for $g \in [0,1]$.

$$\begin{aligned}
\boldsymbol{V}[\hat{\pi}_g|n_{\cdot,g},C,\alpha] &= \boldsymbol{E}_{\boldsymbol{\zeta}_g}[\frac{\pi_g(1-\pi_g)}{n_{\cdot,g}}|n_{\cdot,g}] + \boldsymbol{V}_{\boldsymbol{\zeta}_g}[n_{\cdot,g}\pi_g|n_{\cdot,g}] \\
&= \frac{\boldsymbol{E}_{\boldsymbol{\zeta}_g}[\pi_g]\cdot(1-\boldsymbol{E}_{\boldsymbol{\zeta}_g}[\pi_g])}{n_{\cdot,g}} + \frac{(n_{\cdot,g}-1)}{n_{\cdot,g}}\boldsymbol{V}_{\boldsymbol{\zeta}_g}[\pi_g] \\
&= \frac{\zeta_{1,g}\zeta_{0,g}}{n_{\cdot,g}(\zeta_{1,g}+\zeta_{0,g})^2} + \frac{(n_{\cdot,g}-1)\zeta_{1,g}\zeta_{0,g}}{n_{\cdot,g}(\zeta_{1,g}+\zeta_{0,g})^2(\zeta_{1,g}+\zeta_{0,g}+1)} \\
&= \frac{\zeta_{1,g}\zeta_{0,g}(\zeta_{1,g}+\zeta_{0,g}+n_{\cdot,g})}{n_{\cdot,g}(\zeta_{1,g}+\zeta_{0,g})^2(\zeta_{1,g}+\zeta_{0,g}+1)},
\end{aligned}$$

We can obtain the $100(1-\alpha)\%$ approximate Bayesian credible interval for precision as shown in Eq (6).

Using the same approximation as Katz et al. (1978) (based on the Delta method), the approximate variance of the log ratio $\hat{u}$ is:

$$\boldsymbol{V}[\hat{u}|n_{\cdot,1},n_{\cdot,0},C,\alpha] = \frac{\boldsymbol{V}[\hat{\pi}_1|n_{\cdot,1},C,\alpha]}{(\boldsymbol{E}[\hat{\pi}_1])^2} + \frac{\boldsymbol{V}[\hat{\pi}_0|n_{\cdot,0},C,\alpha]}{(\boldsymbol{E}[\hat{\pi}_0])^2}$$

where $\boldsymbol{E}[\hat{\pi}_g] = \boldsymbol{E}_{\boldsymbol{\zeta}_g}[\boldsymbol{E}[\hat{\pi}_g|\pi_g]] = \boldsymbol{E}_{\boldsymbol{\zeta}_g}[\pi_g] = \frac{\zeta_{1,g}}{\zeta_{1,g}+\zeta_{0,g}}$ for $g \in [0,1]$.

Substituting, we can obtain $\boldsymbol{V}[\hat{u}|n_{\cdot,1},n_{\cdot,0},C,\alpha]$, shown in Eq (8).

Since $\pi_{recall} = 1/(1+\frac{1}{k}\exp(u))$ is a monotonically decreasing function of u, the Bayesian credible interval for u is:

$$\log\frac{\zeta_{1,0}(\zeta_{1,1}+\zeta_{0,1})}{\zeta_{1,1}(\zeta_{1,0}+\zeta_{0,0})} \pm z_{1-\frac{\eta}{2}}\sqrt{\boldsymbol{V}[\hat{u}|n_{\cdot,1},n_{\cdot,0},C,\alpha]}.$$

Similarly to calculating confidence intervals, we can obtain the credible interval for recall, shown in Equation(7).

B PROOFS OF THEOREMS

B.1 Proof of Theorem 1

Proof. Minimizing $\boldsymbol{V}(\pi_{recall})$ is equivalent to minimizing $\boldsymbol{V}(\hat{u})$. We have

$$\boldsymbol{V}(\hat{u}) = \frac{1-\pi_1}{n_{\cdot,1}\pi_1} + \frac{1-\pi_0}{n_{\cdot,0}\pi_0} = \frac{1}{n_{\cdot,1}\Omega_1} + \frac{1}{n_{\cdot,0}\Omega_0} = \frac{1}{v}(\frac{1}{k\cdot s\cdot\Omega_1} + \frac{k\cdot s}{\Omega_0} + \frac{1}{\Omega_1} + \frac{1}{\Omega_0})$$

, where $n_{\cdot 1}$ and $n_{\cdot 0}$ are in Eq (1). Let the derivative of $\boldsymbol{V}(u)$ with respect to u be zero, thus $\boldsymbol{V}(u)$ is minimized at

$$s^* = \frac{1}{k}\sqrt{\Omega_0/\Omega_1} \tag{13}$$

The ratio of the odds Ω_1 and Ω_0 is called the odds ratio. □

B.2 Proof of Theorem 2

Proof. Minimizing $\boldsymbol{V}(\pi_{recall})$ is equivalent to minimizing $\boldsymbol{V}(\hat{u})$. The formula of $\boldsymbol{V}(\hat{u})$ is given in Eq (8). If we remove the terms that do not involve $n_{\cdot,1}$ or $n_{\cdot,0}$ in the formula of $\boldsymbol{V}(\hat{u})$, then minimizing $\boldsymbol{V}(\pi_{recall})$ is equivalent to minimizing the following:

$$\frac{\zeta_{0,1}(\zeta_{1,1}+\zeta_{0,1})}{n_{\cdot,1}\zeta_{1,1}(\zeta_{1,1}+\zeta_{0,1}+1)} + \frac{\zeta_{0,0}(\zeta_{1,0}+\zeta_{0,0})}{n_{\cdot,0}\zeta_{1,0}(\zeta_{1,0}+\zeta_{0,0}+1)} = \frac{1}{n_{\cdot,1}\Theta_1} + \frac{1}{n_{\cdot,0}\Theta_0}$$

where $\Theta_1 = \frac{\zeta_{1,1}(\zeta_{1,1}+\zeta_{0,1}+1)}{\zeta_{0,1}(\zeta_{1,1}+\zeta_{0,1})}$ and $\Theta_0 = \frac{\zeta_{1,0}(\zeta_{1,0}+\zeta_{0,0}+1)}{\zeta_{0,0}(\zeta_{1,0}+\zeta_{0,0})}$. Then, similarly to Theorem 1, $\boldsymbol{V}(\hat{u})$ is minimized at

$$s^* = \frac{1}{k}\sqrt{\Theta_0/\Theta_1} \tag{14}$$

□

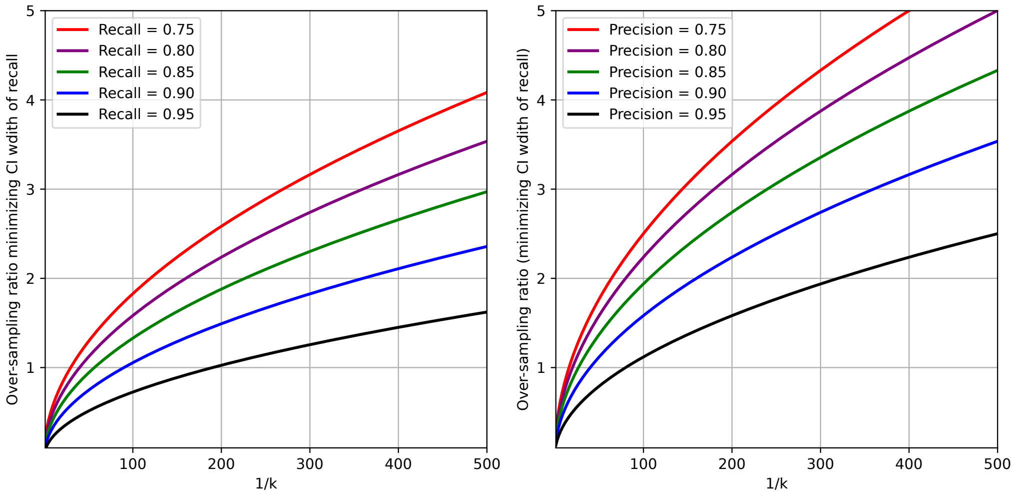

Figure 3: Plots of over-sampling ratio minimizing recall's confidence interval with $1/k$ as the x-axis; Left: plots with various values of recall by fixing $Prec = 0.9$; Right: plots with various values of precision by fixing $Recall = 0.8$.

C OPTIMAL OVER-SAMPLING RATIO s^*

Comparing s^* from the frequentist perspective vs from the Bayesian perspective, when the mean of precision and the mean of recall from the Bayesian posterior distributions match the true precision and recall from the frequentist scenario, the expectations for $Beta(\zeta_{1,1}, \zeta_{0,1})$ and $Beta(\zeta_{1,0}, \zeta_{0,0})$ are equal to π_1 and π_0 respectively. Then we have:

$$\frac{\zeta_{1,1}}{\zeta_{1,1} + \zeta_{0,1}} = \pi_1, \quad \frac{\zeta_{1,0}}{\zeta_{1,0} + \zeta_{0,0}} = \pi_0$$

Correspondingly,

$$\Theta_1 = \frac{\pi_1(\zeta_{1,1} + \zeta_{0,1} + 1)}{(1 - \pi_1)(\zeta_{1,1} + \zeta_{0,1})} \approx \frac{\pi_1}{1 - \pi_1} = \Omega_1 \qquad \Theta_0 = \frac{\pi_0(\zeta_{1,0} + \zeta_{0,0} + 1)}{(1 - \pi_0)(\zeta_{1,0} + \zeta_{0,0})} \approx \frac{\pi_0}{1 - \pi_0} = \Omega_0$$

Thus, the s^* from the frequentist perspective is close to s^* from the Bayesian perspective. To visualize the relationships between k, precision, recall, and s^* (which minimizes the confidence interval width), we plot the s^* (y-axis) and $1/k$ (x-axis) for different values of P/R. See Figure 3. Each curve with fixed P/R has s^* increasing as the data gets more-and-more imbalanced. If we fix k and precision, the lower the recall, the larger s^* needs to be in order to minimize the confidence interval width of recall. If we fix k and recall, the lower the precision, the larger s^* needs to be in order to minimize the confidence interval width of precision. Overall, a higher over-sampling ratio is needed when the data is more imbalanced (smaller k) and P/R is lower.