On Model Evaluation under Non-constant Class Imbalance

Jan Brabec[1,2], Tomáš Komárek[1,2], Vojtěch Franc[2] *, and Lukáš Machlica[3]

[1] Cisco Systems, Inc., Karlovo Namesti 10 Street, Prague, Czech Republic
{janbrabe,tomkomar}@cisco.com
[2] Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Rep.
xfrancv@cmp.felk.cvut.cz
[3] Resistant.ai, Prague, Czech Republic
lukas.machlica@resistant.ai

Abstract. Many real-world classification problems are significantly class-imbalanced to detriment of the class of interest. The standard set of proper evaluation metrics is well-known but the usual assumption is that the test dataset imbalance equals the real-world imbalance. In practice, this assumption is often broken for various reasons. The reported results are then often too optimistic and may lead to wrong conclusions about industrial impact and suitability of proposed techniques. We introduce methods[4] focusing on evaluation under non-constant class imbalance. We show that not only the absolute values of commonly used metrics, but even the order of classifiers in relation to the evaluation metric used is affected by the change of the imbalance rate. Finally, we demonstrate that using subsampling in order to get a test dataset with class imbalance equal to the one observed in the wild is not necessary, and eventually can lead to significant errors in classifier's performance estimate.

Keywords: Evaluation metrics · Imbalanced data · Precision · ROC

1 Introduction

Class-imbalanced problems arise if number of samples in one of the classes, often in the class of interest, is significantly lower than in the other class, often the background class. Such problems are present in variety of different domains such as medicine [16], finance [15, 20, 21], cybersecurity [1, 3, 5] and many others.

In highly imbalanced problems it is essential to use suitable evaluation metrics to correctly assess the merit of pursued algorithms and realistically judge their impact before they are deployed into the wild. Methods for evaluation of classifiers on class-imbalanced datasets are well known and have been thoroughly described in the past [4, 9, 11, 19].

* VF was supported by OP VVV project CZ.02.1.01\0.0\0.0\16_019\0000765 Research Center for Informatics.

[4] Supplementary code related to techniques described in this paper is available at: https://github.com/CiscoCTA/nci_eval

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

It is usually assumed that the imbalance of the test dataset is the same as in the real distribution on which the model will operate once deployed into production environment. However, this assumption is often broken, because of different reasons ranging from selection bias when constructing the test dataset, high costs of acquiring large dataset mainly in situations when the imbalance is high (e.g. $1 : 10^4$), to the fact that often not a single general distribution exists (e.g. disease classifier may face different priors depending on the location).

Discrepancy between imbalances in test datasets and real world is often the root cause of too optimistic results leading to wrong expectations of the impact in industrial applications. This is detrimental to the research community, because it creates confusion about which problems are still open and which are solved. It might discourage groups from working on such problems, and make it harder for researchers still investigating the field to convince the community that in the light of the too optimistic prior work their results have still impact.

Throughout this paper, we frame and investigate the problem of classifier evaluation dropping the assumption of constant class imbalance.We focus on precision related metrics as one of the most popular metrics for imbalanced problems [4, 9]. We show how these metrics can be computed for arbitrary class imbalances and any test dataset without the need to re-sample the data. We also inspect their behavior as a function of the imbalance rate. We show that Precision-Recall (PR) curves have little value without stating the corresponding imbalance ratio which can dramatically affect the results and their assessment.

We demonstrate that change in imbalance rate, maybe surprisingly, affects also the ranking of classifiers under these metrics. We argue that instead of tabulating the results for a single dataset, it is beneficial to plot the dependence on the class imbalance rate whenever possible. Such plots provide considerably more information for wider audience.

We also describe how errors in measurements can be assessed and that they can significantly affect the reliability of measured precision mainly in cases when low regions of false positive rate are of interest. This can be primarily attributed to the fact that the test dataset is finite. Therefore, we further elaborate how the class imbalance increases the demands on the size of test dataset.

Most importantly, *we refute the common understanding that the best practice is to alter the test dataset so that class imbalance matches the imbalance of the pursued distribution* as is suggested e.g. in [14]. We show how re-sampling of a dataset may lead to significant errors in measurements. We stress that the test dataset should be constructed in a way to allow measurements of false-positive and true-positive rates with errors as small as possible. We show that the crucial entity to focus on is the coefficient of variation related to both true-positive and false-positive rates.

2 Preliminaries

Throughout this paper we are concerned with the binary classification task. Let $\boldsymbol{x} \in \mathcal{X}$ be an input and $y \in \mathcal{Y} = \{-1, 1\}$ be a target. We call the class $y = -1$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

negative class and the class $y = 1$ positive class. The positive class is assumed to be the minority class and the negative class is the majority class. We do not assume that there exists a single real-world joint-probability distribution $p(\boldsymbol{x}, y)$ but instead consider a parametric family:

$$p(\boldsymbol{x}, y; \eta) = p(\boldsymbol{x}|y) \cdot P(y; \eta), \text{ where } P(y; \eta) = \begin{cases} 1 - \eta & y = -1 \\ \eta & y = 1 \end{cases}. \tag{1}$$

Parameter $\eta \in [0, 1]$ specifies the positive class prevalence. If we consider a classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ then the following classifier evaluation metrics can be expressed as probabilities:

$$\text{TPR} = \text{Recall} = P(h(\boldsymbol{x}) = 1|y = 1) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|y=1)}[h(\boldsymbol{x}) = 1] \tag{2}$$

$$\text{FPR} = P(h(\boldsymbol{x}) = 1|y = -1) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|y=-1)}[h(\boldsymbol{x}) = 1] \tag{3}$$

$$\text{Prec}(\eta) = P(y = 1|h(\boldsymbol{x}) = 1) = \frac{\text{TPR} \cdot \eta}{\text{TPR} \cdot \eta + \text{FPR} \cdot (1 - \eta)} \tag{4}$$

TPR stands for true-positive-rate (also called recall or sensitivity), FPR for false-positive-rate and Prec for precision. Formula (4) is derived using Bayes' theorem. We can observe that both TPR and FPR are not affected by the positive class prevalence but precision is. This observation is very important for the rest of this paper.

To estimate the above-mentioned metrics we need to evaluate the classifier on a test dataset. We assume that the test dataset is sampled i.i.d. from $p(\boldsymbol{x}, y; \eta_{test})$ where η_{test} may or may not correspond to a positive class prevalence connected to some real-world application of the classifier. TP, FP, TN, FN denote the number of true positives, false positives, true negatives and false negatives, respectively and $N = TP + FP + TN + FN$ equals the size of the test set.

Prevalence of the positive class in the test dataset p_+ and imbalance ratio (IR) are defined as (one can be computed from the other easily):

$$p_+ = \frac{TP + FN}{N}, \quad IR = \frac{TP + FN}{TN + FP}. \tag{5}$$

$\widehat{\text{TPR}}$ is defined as the fraction of positive samples that were classified correctly:

$$\widehat{\text{TPR}} = \frac{1}{|\mathcal{X}^+|} \sum_{\boldsymbol{x} \in \mathcal{X}^+} [\![h(\boldsymbol{x}) = 1]\!] = \frac{\text{TP}}{|\mathcal{X}^+|} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

where $[\![\cdot]\!]$ is the indicator function. $\widehat{\text{FPR}}$ is defined as the fraction of negatives samples that were classified incorrectly:

$$\widehat{\text{FPR}} = \frac{1}{|\mathcal{X}^-|} \sum_{\boldsymbol{x} \in \mathcal{X}^-} [\![h(\boldsymbol{x}) = 1]\!] = \frac{\text{FP}}{|\mathcal{X}^-|} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{7}$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

$\widehat{\text{Prec}}$ is the number of true positives out of all the positive predictions:

$$\widehat{\text{Prec}}(\eta) = \frac{\widehat{\text{TPR}} \cdot \eta}{\widehat{\text{TPR}} \cdot \eta + \widehat{\text{FPR}} \cdot (1 - \eta)} \tag{8}$$

It can be easily shown that $\widehat{\text{Prec}}(p_+) = \text{TP}/(\text{TP} + \text{FP})$ resolves to the standard formula used to compute precision. It holds that the metrics measured on the test dataset approach their true values originating from the distribution $p(\boldsymbol{x}, y; \eta)$ as the size of the dataset grows. In other words $p_+ \to \eta_{test}$, $\widehat{\text{TPR}} \to \text{TPR}$, $\widehat{\text{FPR}} \to \text{FPR}$ and $\widehat{\text{Prec}} \to \text{Prec}$ as N approaches infinity, but the errors in estimation caused by limited size of test dataset are often significant enough to deserve consideration, particularly during classifier evaluation in settings that are heavily class-imbalanced. We elaborate on this in Section 5.

3 Precision in the light of different class imbalance ratios

Equation (8) in Section 2 shows that the class imbalance ratio of the test dataset directly impacts the measured precision. As such, the test dataset class imbalance must be considered when interpreting the results to assess viability of the classifier for a given application.

Fortunately, it is not necessary for a test dataset's imbalance ratio to be equivalent to the real-world imbalance. Equation (8) shows how to estimate precision ($\widehat{\text{Prec}}$), that corresponds to any class imbalance, from $\widehat{\text{TPR}}$ and $\widehat{\text{FPR}}$ which are estimated from the test dataset and are unaffected by it's imbalance.

In Section 5 we provide rationale and show that matching the real-world class imbalance is often sub-optimal and not desirable for correct evaluation.

3.1 Positive-prevalence precision curve

Positive prevalence adjusted precision computed by Equation (8) is a linear rational function of the positive class prevalence η. As such, it can be plotted over an interval of positive prevalence values. We call such plot Positive-Prevalence Precision (P^3) curve. The curve should be plotted with log-scaled x-axis (lin-log P^3 curve) to easily distinguish between different orders of magnitude of the positive prevalence as demonstrated in Figure 1.

P^3 curve is a useful instrument when evaluating a classifier to determine it's performance beyond a particular dataset. The downside of the plot is that contrary to ROC or PR curves, it captures the performance only for a single operating point of the classifier. Each point on an ROC curve thus has it's corresponding P^3 curve.

Given a particular ROC curve, each point on the curve corresponds to a different value of TPR. Instead of saying that P^3 curve corresponds to a particular point on the ROC curve, it can also be said that it corresponds to a fixed value of TPR. For example, P^3 curve in Figure 1 corresponds to a classifier with TPR fixed at 60%.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

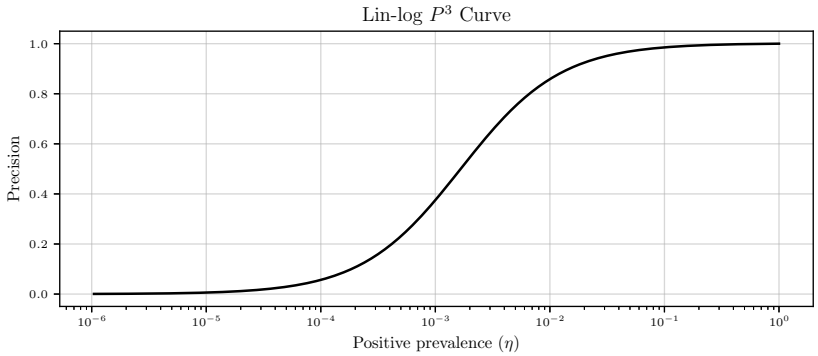

Fig. 1. Positive-Prevalence Precision (P^3) curve for a hypothetical classifier with TPR = 0.6 and FPR = 0.001. The graph is plotted in logarithmic scale of the x-axis.

P^3 curve answers the question "How does precision of a given classifier evolve when changing the class imbalance-ratio?" and allows to quickly visually assess some of the conditions under which the classifier is suitable for production environment. Also, even if P^3 curve may not be used in a particular evaluation of a classifier it is still important to possess intuition about it's general shape.

3.2 Precision-Recall curves

PR curve is a very popular method to evaluate classifiers on imbalanced datasets. It captures the relationship between recall (TPR) on the x-axis and precision on the y-axis. As is the case with ROC curve, PR curve is usually created by applying different thresholds on the raw output of a classifier. While ROC is a strictly increasing function, PR curves do not have to be monotonous because it is possible for precision to both increase or decrease for different threshold values.

As discussed in Section 3.1, contrary to the ROC curve, *PR curve is affected by the imbalance ratio present in the test dataset.* This behavior is demonstrated in Figure 2. PR curves can immediately reveal poor performance on class-imbalanced datasets that might not be obvious when inspecting ROC curves alone [18]. Because of this property PR curves are well suited and popular choice for evaluation of classifiers on class-imbalanced sets.

We suggest that the particular imbalance ratio present in a test dataset for which the PR curve was created should always be reported and considered when interpreting the impact of the results. When different research teams perform their experiments on different test sets while solving the same problem, and even if the data originate from the same source, the resulting PR curves will not be comparable if different imbalance ratios are present. For example, in computer security the datasets of downloaded files might originate from the VirusTotal[5]

[5] https://www.virustotal.com

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

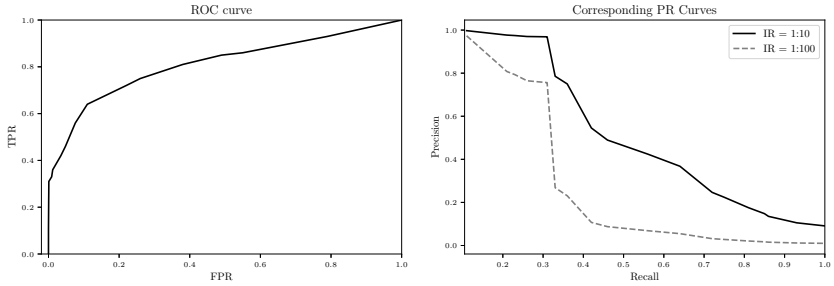

Fig. 2. Example of how a single ROC curve can correspond to two different PR curves given different imbalance ratios. The solid PR curve was created from the ROC curve with assumption that the IR was $1 : 10$ while the dashed PR curve corresponds to IR equal to $1 : 100$.

service, but different teams may work with different subsets that have different imbalance ratios.

Another danger is that the class imbalance ratio in a particular test dataset is often not representative of the imbalance ratios encountered once the classifier is deployed in real environment. It is often the case that the imbalance ratios experienced in the wild are lower than the ratio in the test dataset (not rarely the test datasets are even not imbalanced at all). In such situations, too optimistic estimates of the classifier's performance will be obtained if evaluation based on PR curve computed directly on the test dataset is used.

To remedy these risks, often test datasets with the same class imbalance ratios that would be encountered in the real environment are created. In Section 5 we demonstrate that this should not be the goal. Rather, a test dataset should be assembled that allows estimation of TPR and FPR with low enough variance and (8) should be used to compute Precision-Recall curves for different class imbalance ratios of interest.

4 Comparing performance of classifiers

When comparing performance of classifiers that need to deal with imbalanced data, the area under PR-curve (PR-AUC) or F1 score ($F_1 = 2 \cdot \frac{\text{Prec} \cdot \text{Recall}}{\text{Prec} + \text{Recall}}$) are often used out of convenience because they can be expressed as a single number [8]. In this section, we show that not only the values of these metrics dramatically depend on the imbalance rate in the selected test dataset, but the rate has notable influence even on the order of classifiers related to their efficacy. That is, based on these metrics two classifiers can switch places given different imbalance rates. This can lead to incorrect conclusions about performance of classifiers on real data. The fact can be also misused for cherry-picking of an imbalance rate to pick the one where a classifier achieves better results than any other method it competes with.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

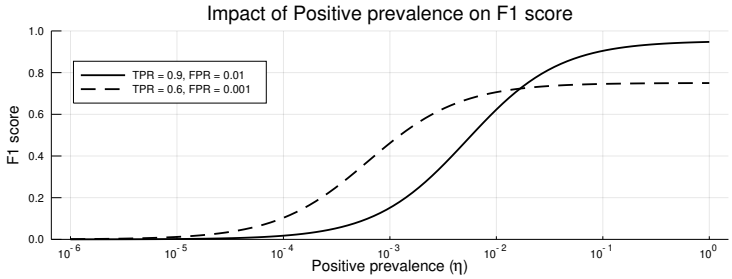

Fig. 3. The graph is similar to Positive Prevalence-Precision plot in Figure 1 but instead of precision it plots F1 score of two distinct classifiers computed on the same dataset but assuming different imbalance rates. It can be seen that not only the absolute value of the score but even the order of the classifiers depends on the positive class prevalence.

4.1 Affecting ordering of classifiers: F1 score

F1 score is defined as harmonic mean of precision and recall. The comparison of F1 scores of two classifiers is therefore affected by the selected imbalanced rate since precision depends on the rate while recall does not. Figure 3 demonstrates how the F1 score of two classifiers depends on the imbalance rate present in a test dataset.

Therefore, *we suggest to plot F1 scores in relation to imbalance rates, such as seen in Figure 3 instead of tabulated F1 scores in any applied research papers.* The plot contains a superset of information, it is easily interpretable, space-efficient and conveys an overall better picture about performance of classifiers independent of the particular imbalance rate in the selected test dataset. The imbalance rate of the particular test dataset can be easily highlighted on the x-axis.

4.2 Affecting ordering of classifiers: PR-AUC

Firstly, it is proven that if a classifier *dominates* in ROC space it also dominates in PR space [6], but dominance is not linked to the area under ROC curve (ROC-AUC). It is easily possible for a classifier to have greater ROC-AUC than another but smaller area under PR curve (PR-AUC) on the same test dataset.

A convenient property of evaluating classifier by ROC-AUC is that it's value is invariant to class imbalance. On the other hand, the value of ROC-AUC can be dominated by insignificant regions in the ROC space, e.g. high values of FPR, which are in practice of no importance. If the problem is heavily class imbalanced it is usually not an appropriate method for evaluation of classifiers [2] and PR-AUC should be considered.

However, it is often not realized that PR-AUC values depend on class imbalance and notably that also the order of classifiers under this metric depends on the imbalance rate as demonstrated in Figure 4. It may be more surprising

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

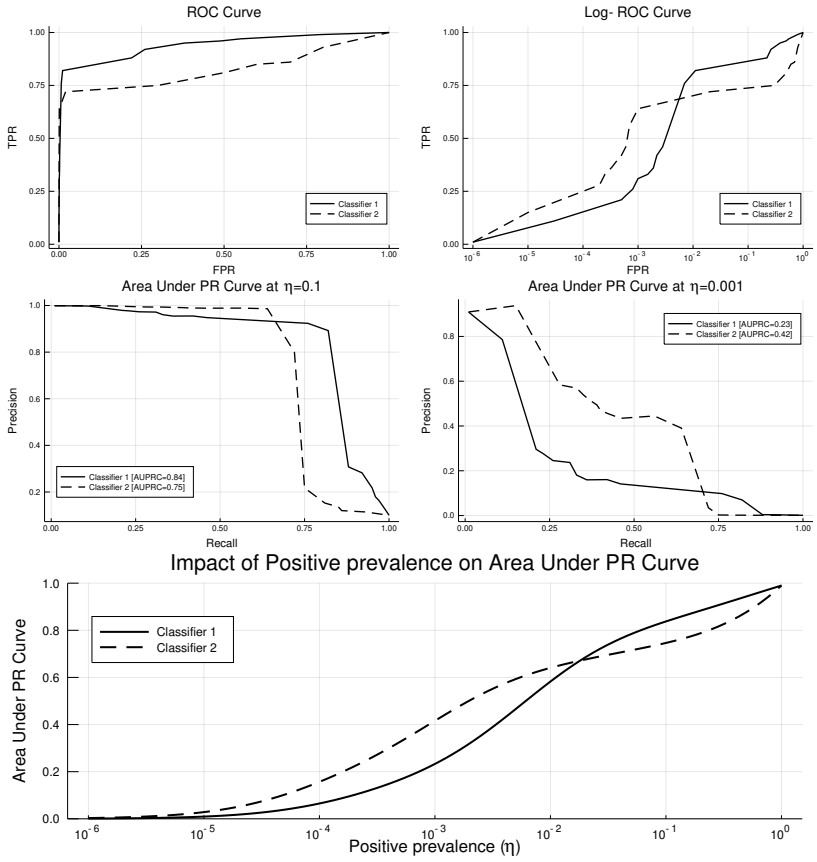

Fig. 4. The top-left plot is an example plot of two classifier ROC curves. In the top-right plot the same ROC curves are displayed with logarithmically scaled x-axis. The middle row displays corresponding PR curves for the ROC curves under different positive class prevalences (namely 10^{-1} and 10^{-3}). The bottom plot shows how PR-AUC of the classifiers depends on the class imbalance rate and that the order of the classifiers can easily switch for two different prevalences.

than in the case of F1 score computed only at a single operating point, while PR-AUC is evaluated over the whole range of operating points. Therefore, one might wrongly expect the metric to preserve ordering of classifiers across different imbalance rates.

We offer similar advice as with F1 score about the need to report the dataset imbalance rate together with PR-AUC values and to ideally use plots as in Figure 4 instead of tabulated values for a single imbalance ratio.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

5 Impact of errors on estimates of TPR and FPR

Class-imbalanced problems have increased demands on the test dataset size. It is often ignored that $\widehat{\mathrm{TPR}}$ and $\widehat{\mathrm{FPR}}$ computed on test dataset are just point estimates of the real TPR and FPR, given in (2) and (3), respectively, and as such they may be affected by uncertainty related to insufficient amount of samples of the minority class. In this section, we investigate how this uncertainty impacts the measured precision and how to correctly design experiments in presence of imbalanced data to suppress the uncertainty in the outcome.

A common approach to quantify the uncertainty of estimates based on finite samples is to use the interval estimates. We say that $\mathcal{I}_{\mathrm{TPR}} = (\widehat{\mathrm{TPR}} - \sigma_{\mathrm{TPR}}, \widehat{\mathrm{TPR}} + \sigma_{\mathrm{TPR}})$ is the α-confidence interval of TPR if it holds that

$$\mathrm{Prob}(\mathrm{TPR} \in \mathcal{I}_{\mathrm{TPR}}) \geq \alpha\,, \tag{9}$$

where the probability is w.r.t. randomly generated positive test samples $\mathcal{X}^+$ which are used to compute $\widehat{\mathrm{TPR}}$ by (6). The interval (half-)width σ_{TPR}, the number of samples $|\mathcal{X}^+|$ and the confidence level $\alpha \in (0,1)$ are dependent variables the exact relation of which is characterized by numerous concentration bounds like the Hoeffding's inequality. For example, by fixing σ_{TPR} and α we can compute the minimal number of samples in $\mathcal{X}^+$ which guarantee that $\mathcal{I}_{\mathrm{TPR}}$ is the α-confidence interval. In the sequel we assume that the interval width σ_{TPR} is not greater than $\widehat{\mathrm{TPR}}$. Note that this formalisation does not introduce any specific constraints on the shape of TPR distribution. The confidence interval $\mathcal{I}_{\mathrm{TPR}}$ can be characterized by a single number, the coefficient of variation, defined as

$$\mathrm{CV}_{\mathrm{TPR}} = \frac{\sigma_{\mathrm{TPR}}}{\widehat{\mathrm{TPR}}}\,. \tag{10}$$

Analogously, we can define $\mathcal{I}_{\mathrm{FPR}} = (\widehat{\mathrm{FPR}} - \sigma_{\mathrm{FPR}}, \widehat{\mathrm{FPR}} + \sigma_{\mathrm{FPR}})$, $\mathrm{CV}_{\mathrm{FPR}} = \frac{\sigma_{\mathrm{FPR}}}{\widehat{\mathrm{FPR}}}$, and we also assume that $\sigma_{\mathrm{FPR}} < \widehat{\mathrm{FPR}}$.

Let us define the precision as a function of the positive class prevalence η, TPR and FPR [6]:

$$\mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR}) = \frac{\eta \cdot \mathrm{TPR}}{\eta \cdot \mathrm{TPR} + (1-\eta) \cdot \mathrm{FPR}}\,. \tag{11}$$

Given $\mathrm{TPR} \in \mathcal{I}_{\mathrm{TPR}}$ and $\mathrm{FPR} \in \mathcal{I}_{\mathrm{FPR}}$, the value of $\mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR})$ has to be for any fixed $\eta \in (0,1)$ inside the interval $(\mathrm{LB}(\eta), \mathrm{UB}(\eta))$ where

$$\mathrm{LB}(\eta) = \min_{\substack{\mathrm{TPR} \in \mathcal{I}_{\mathrm{TPR}} \\ \mathrm{FPR} \in \mathcal{I}_{\mathrm{FPR}}}} \mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR})\,, \tag{12}$$

$$\mathrm{UB}(\eta) = \max_{\substack{\mathrm{TPR} \in \mathcal{I}_{\mathrm{TPR}} \\ \mathrm{FPR} \in \mathcal{I}_{\mathrm{FPR}}}} \mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR})\,. \tag{13}$$

Let Δ be the maximal width of the interval $(\mathrm{LB}(\eta), \mathrm{UB}(\eta))$ w.r.t. η, that is,

$$\Delta = \max_{\eta \in (0,1)} (\mathrm{UB}(\eta) - \mathrm{LB}(\eta))\,. \tag{14}$$

[6] In (8) we used $\mathrm{Prec}(\eta)$ since the values of TPR and FPR were assumed to be fixed.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

The number Δ can be interpreted as the maximal uncertainty in measurements of precision when the exact values of TPR and FPR are replaced by their confidence intervals $\mathcal{I}_{\mathrm{TPR}}$ and $\mathcal{I}_{\mathrm{FPR}}$, respectively. It is easy to see that $\mathrm{TPR} \in \mathcal{I}_{\mathrm{TPR}}$ and $\mathrm{FPR} \in \mathcal{I}_{\mathrm{FPR}}$ imply

$$\mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR}) \in (\widehat{\mathrm{Prec}}(\eta) - \Delta, \widehat{\mathrm{Prec}}(\eta) + \Delta)\,. \tag{15}$$

The concepts of $\mathrm{UB}(\eta)$, $\mathrm{LB}(\eta)$ and Δ as well as their relation to $\mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR})$ are illustrated in Figure 5. The following theorem relates the maximal uncertainty Δ and the coefficients of variation $\mathrm{CV}_{\mathrm{TPR}}$ and $\mathrm{CV}_{\mathrm{FPR}}$, which characterize the confidence intervals $\mathcal{I}_{\mathrm{TPR}}$ and $\mathcal{I}_{\mathrm{FPR}}$, respectively.

Theorem 1. *Let* $TPR \in (\widehat{TPR} - \sigma_{TPR}, \widehat{TPR} + \sigma_{TPR})$ *and* $FPR \in (\widehat{FPR} - \sigma_{FPR}, \widehat{FPR} + \sigma_{FPR})$. *Let further* $\widehat{TPR} > \sigma_{TPR}$ *and* $\widehat{FPR} > \sigma_{FPR}$. *Then*

$$\Delta \leq \max\{CV_{TPR}, CV_{FPR}\}$$

and the equality is attained iff $CV_{TPR} = CV_{FPR}$.[7]

Corollary 1. *Let* $\mathcal{I}_{TPR}$ *and* $\mathcal{I}_{FPR}$ *be* α*-confidence intervals of the true TPR and FPR, respectively, and let* CV_{TPR} *and* CV_{FPR} *be their corresponding coefficients of variation. Let further* $\Delta = \max\{CV_{TPR}, CV_{FPR}\}$. *Then* $\mathcal{I}_{Prec} = (\widehat{Prec}(\eta) - \Delta, \widehat{Prec}(\eta) + \Delta)$ *is the* α^2*-confidence interval of* $Prec(\eta, TPR, FPR)$, *i.e.*

$$Prec(\eta, TPR, FPR) \in \mathcal{I}_{Prec}$$

holds with probability α^2 *at least.*

The α^2-confidence level stems from the fact that $\mathrm{TPR} \in \mathcal{I}_{\mathrm{TPR}}$ and $\mathrm{FPR} \in \mathcal{I}_{\mathrm{FPR}}$ are two independent random events with probability not less than α.

Theorem 1 shows the relationship between confidence intervals for precision, widths of these intervals and point estimates of TPR, FPR. That is, coefficients of variation for TPR and FPR are the crucial quantities to consider when designing test dataset. If a test set is constructed we first need to manually fix both σ_{TPR} and σ_{FPR} at reasonable values based on the purpose of the dataset, and then ensure sufficient number of testing samples necessary to estimate TPR, FPR with desired Δ. If, for example, one is interested in $\mathrm{FPR} = 10^{-3}$ on a dataset having only 10,000 negative samples, the estimate around this working point may become extremely noisy. Since such low FPR corresponds to only 10 FP samples ($10{,}000 * 10^{-3}$), just a small increase or decrease in number of FPs suffice to significantly alter the relative value of the FPR. Therefore, if such low values of FPR are of interest, one should increase the amount of negatives. Different methods exist that can quantify the concentration bounds. For example, Hoeffding's inequality can be used, which states that the upper bound on the number of required samples is proportional to $\frac{1}{\sigma_{\mathrm{FPR}}{}^2}$, but Hoeffding's bound is very loose and usually less samples are required.

[7] The proof for Theorem 1 is available in the appendix of this paper at: https://arxiv.org/pdf/2001.05571.pdf

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

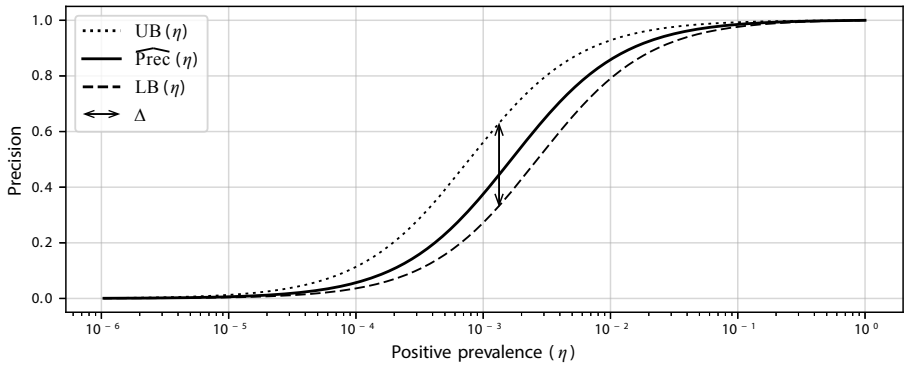

Fig. 5. The figure visualizes the uncertainty band containing the value of $\mathrm{Prec}(\eta, \mathrm{TPR}, \mathrm{FPR}) \in (\mathrm{UB}(\eta), \mathrm{LB}(\eta))$ when TPR and FPR are bound to intervals $\mathcal{I}_{\mathrm{TPR}} = (\widehat{\mathrm{TPR}} - \sigma_{\mathrm{TPR}}, \widehat{\mathrm{TPR}} + \sigma_{\mathrm{TPR}})$ and $\mathcal{I}_{\mathrm{FPR}} = (\widehat{\mathrm{FPR}} - \sigma_{\mathrm{FPR}}, \widehat{\mathrm{FPR}} + \sigma_{\mathrm{FPR}})$, respectively. The value $\Delta = \max_{\eta \in (0,1)}(\mathrm{UB}(\eta) - \mathrm{LB}(\eta))$ corresponds to the maximal width of the uncertainty band. The solid line corresponds to the point estimate $\widehat{\mathrm{Prec}}(\eta) = \mathrm{Prec}(\eta, \widehat{\mathrm{TPR}}, \widehat{\mathrm{FPR}})$.

On the other hand, given a test dataset, in order to find Δ we need to estimate $\sigma_{\mathrm{TPR}}, \sigma_{\mathrm{FPR}}$ to get $\mathrm{CV}_{\mathrm{TPR}}, \mathrm{CV}_{\mathrm{FPR}}$. For that purpose cross-validation or bootstrapping can be used. For example, a classifier with $\widehat{\mathrm{TPR}} = 0.6, \sigma_{\mathrm{TPR}} = 0.06, \widehat{\mathrm{FPR}} = 10^{-3}, \sigma_{\mathrm{FPR}} = 10^{-4}$ has $\mathrm{CV}_{\mathrm{TPR}} = \mathrm{CV}_{\mathrm{FPR}} = \Delta = 0.1$, which might be reasonable width of the precision's confidence interval (i.e. $\pm 10\%$ change). But, if we increase $\sigma_{\mathrm{FPR}} = 5 * 10^{-4}$ then even though the number might seem small and it may be not indicative of the impact on estimate of the precision, the bound for precision becomes $\Delta = 0.5$ (i.e. $\pm 50\%$ change), which will immediately shed light on the reliability of estimates of the precision.[8]

5.1 Example of errors caused by sub-sampling

To illustrate the error of sub-sampling we used ResNet-50 [10] on the ImageNet validation dataset [17] to detect images of 'agama' in a one-vs-all manner. The p_+ in such dataset is 10^{-3}.

To plot PR curves for $\eta = 10^{-2}$ we can either use the full dataset and then apply (8) to adjust the precision, or sub-sample the dataset to $p_+ = 10^{-2}$. Figure 6 compares these two approaches, where we repeated the sub-sampling 30 times to estimate the variance introduced by random reduction of the negative class. The results show that PR curves measured on the sub-sampled datasets are encumbered by a considerable measurement errors even though each one has 5000 samples, which might otherwise be a reasonable number for evaluation on balanced problems. Moreover $\eta = 10^{-2}$ is not as drastic imbalance as is often

[8] In this example, $\Delta \approx 0.31$ for $\eta \approx 1.45 \cdot 10^{-3}$. Computation can be found in the supplementary code to this paper.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

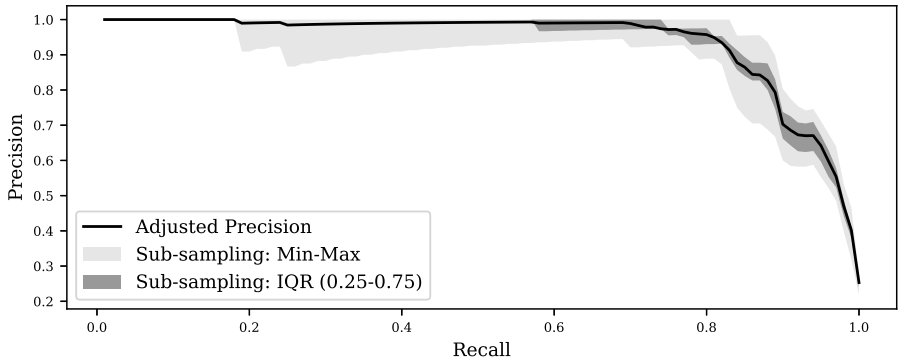

Fig. 6. PR curves for $\eta = 10^{-2}$. The black PR curve is computed from full dataset with $p_+ = 10^{-3}$ and adjusted to $\eta = 10^{-2}$ using (8), whereas the gray areas indicate IQR and min-max range of PR curves computed on 30 datasets with randomly sub-sampled negative class to match $p_+ = 10^{-2}$. Note that some PR curves are inside of IQR only partially.

encountered in applications and the errors could be even more pronounced if η was lower.

Unlike the common practice of sub-sampling of the test dataset to the desired imbalance rate [14], we recommend to use a bigger dataset (to decrease the coefficients of variation) and adjust the metrics to the desired imbalance rate instead.

6 Related Work

Several comprehensive papers about methodology of evaluation on imbalanced datasets were written [4, 7, 9, 11, 19]. They focus on measuring the performance on the test dataset and do not address the problem of mismatch between class imbalances in test and application datasets.

In [5] authors use a plot with area under PR curve on the y-axis and a quantity related to the imbalance ratio on the x-axis. The plot is similar to Figure 4, it is used because it is useful in the context of the paper but it's properties and impacts are not discussed.

In [2] authors discuss several bad practices in handling of class-imbalanced problems. Apart from other causes, they discuss the importance of addressing the real imbalance ratios that can be different from the test dataset. They also present a formula for adjusting the precision to different imbalance ratios but do not explore this formula in greater detail neither inspect the impact of uncertainty originating from the finite size of the test dataset on precision.

Paper [12] introduces measure based on area under PR curve, which is further integrated across different class imbalances yielding a single evaluation number. The idea is based on the relationship between PR and ROC given in (8). No

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

additional investigations related to multiple working points, ordering of classifiers according to the score, nor errors in measurements are carried out.

In [14] authors raise the issue of experimental results in cybersecurity often not being reproducible in real applications. They mention the problem that the class imbalance is often different in test dataset and in practice. They do not address the issue analytically but instead choose to re-sample the test dataset to desired imbalance ratios. This goes directly against our observations in Section 5 and applying such method leads to results heavily affected by noise.

It should be mentioned that other evaluation metrics well-suited for evaluation of class-imbalanced problems were proposed. A notable example is Matthews Correlation Coefficient (MCC) [13], but is not in the scope of this paper. MCC is not as widely used as PR [8] and it's values are not that easily interpretable as values of precision and recall.

7 Conclusion

This paper addressed evaluation of classifiers under consideration that the class imbalance ratio encountered in real world is different from imbalance present in the test dataset or is suspect to change. We focused on precision as one of the most popular evaluation metrics for imbalanced problems.

We stress that it is of significant importance to report also the imbalance ratio under which the classifier was developed and is aimed for, because assuming different imbalance ratios may easily lead to swapping of places of classifiers. This holds also for both PR-AUC and F1 score.

We have shown that even very small absolute values of σ_{FPR} can result in large variance in measured precision. The larger the class imbalance, the greater are the demands on the amount of negative samples present in the test dataset. Therefore, rather than sub-sampling a dataset to reach desired imbalance rate, all the samples should be kept to decrease the coefficients of variation, and the evaluation metrics should be computed given the presented formulas.

References

1. Axelsson, S., Sands, D.: The base-rate fallacy and the difficulty of intrusion detection. Understanding Intrusion Detection Through Visualization pp. 31–47 (2006)
2. Brabec, J., Machlica, L.: Bad practices in evaluation methodology relevant to class-imbalanced problems. Critiquing and Correcting Trends in Machine Learning workshop at NeurIPS **abs/1812.01388** (2018), http://arxiv.org/abs/1812.01388
3. Brabec, J., Machlica, L.: Decision-forest voting scheme for classification of rare classes in network intrusion detection. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3325–3330 (Oct 2018). https://doi.org/10.1109/SMC.2018.00563
4. Chawla, N.V.: Data mining for imbalanced datasets: An overview. In: Data mining and knowledge discovery handbook, pp. 875–886. Springer (2009)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6

5. Damodaran, A., Di Troia, F., Visaggio, C.A., Austin, T.H., Stamp, M.: A comparison of static, dynamic, and hybrid analysis for malware detection. Journal of Computer Virology and Hacking Techniques **13**(1), 1–12 (2017)
6. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240. ACM (2006)
7. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)
8. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
9. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
11. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al.: Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering **30**(1), 25–36 (2006)
12. Landgrebe, T.C., Paclik, P., Duin, R.P.: Precision-recall operating characteristic (p-roc) curves in imprecise environments. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. vol. 4, pp. 123–127. IEEE (2006)
13. Matthews, B.W.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure **405**(2), 442–451 (1975)
14. Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., Cavallaro, L.: {TESSERACT}: Eliminating experimental bias in malware classification across space and time. In: 28th {USENIX} Security Symposium ({USENIX} Security 19). pp. 729–746 (2019)
15. Phua, C., Alahakoon, D., Lee, V.: Minority report in fraud detection: Classification of skewed data. SIGKDD Explor. Newsl. **6**(1), 50–59 (Jun 2004). https://doi.org/10.1145/1007730.1007738, http://doi.acm.org/10.1145/1007730.1007738
16. Rahman, M.M., Davis, D.: Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing **3**(2), 224 (2013)
17. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
18. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), e0118432 (2015)
19. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009)
20. Wei, W., Li, J., Cao, L., Ou, Y., Chen, J.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web **16**(4), 449–475 (2013)
21. Yu, L., Wang, S., Lai, K.K., Wen, F.: A multiscale neural network learning paradigm for financial crisis forecasting. Neurocomputing **73**(4-6), 716–725 (2010)

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_6