Recovering True Classifier Performance in Positive-Unlabeled Learning

Shantanu Jain, Martha White, Predrag Radivojac
Department of Computer Science
Indiana University, Bloomington, Indiana, USA
{shajain, martha, predrag}@indiana.edu

Abstract

A common approach in positive-unlabeled learning is to train a classification model between labeled and unlabeled data. This strategy is in fact known to give an optimal classifier under mild conditions; however, it results in biased empirical estimates of the classifier performance. In this work, we show that the typically used performance measures such as the receiver operating characteristic curve, or the precision-recall curve obtained on such data can be corrected with the knowledge of class priors; i.e., the proportions of the positive and negative examples in the unlabeled data. We extend the results to a noisy setting where some of the examples labeled positive are in fact negative and show that the correction also requires the knowledge of the proportion of noisy examples in the labeled positives. Using state-of-the-art algorithms to estimate the positive class prior and the proportion of noise, we experimentally evaluate two correction approaches and demonstrate their efficacy on real-life data.

Introduction

Performance estimation in binary classification is tightly related to the nature of the classification task. As a result, different performance measures may be directly optimized during training. When (mis)classification costs are available, the classifier is ideally trained and evaluated in a cost-sensitive mode to minimize the expected cost (Whalen 1971; Elkan 2001). More often, however, classification costs are unknown and the overall performance is assessed by averaging the performance over a range of classification modes. The most extensively studied and widely used performance evaluation in binary classification involves estimating the Receiver Operating Characteristic (ROC) curve that plots the true positive rate of a classifier as a function of its false positive rate (Fawcett 2006). The ROC curve provides insight into trade-offs between the classifier's accuracies on positive versus negative examples over a range of decision thresholds. Furthermore, the area under the ROC curve (AUC) has a meaningful probabilistic interpretation that correlates with the ability of the classifier to separate classes and is often used to rank classifiers (Hanley and McNeil 1982). Another important performance criterion generally used in information retrieval relies on the precision-recall (pr-rc) curve, a plot of precision as a function of recall. The precision-recall evaluation, including summary statistics derived from the pr-rc curve, may be preferred to ROC curves when classes are heavily skewed (Davis and Goadrich 2006).

Although model learning and performance evaluation in a supervised setting are well understood (Hastie et al. 2001), the availability of unlabeled data gives additional options and also presents new challenges. A typical semi-supervised scenario involves the availability of positive, negative and (large quantities of) unlabeled data. Here, the unlabeled data can be used to improve training (Blum and Mitchell 1998) or unbias the labeled data (Cortes et al. 2008); e.g., to estimate class proportions that are necessary to calibrate the model and accurately estimate precision when class balances (but not class-conditional distributions) in labeled data are not representative (Saerens et al. 2002). This is often the case when it is more expensive or difficult to label examples of one class than the examples of the other. A special case of the semi-supervised setting arises when the examples of only one class are labeled. It includes open-world domains such as molecular biology where, for example, wet lab experiments determining a protein's activity are generally conclusive; however, the absence of evidence about a protein's function cannot be interpreted as the evidence of absence. This is because, even when the labeling is attempted, a functional assay may not lead to the desired activity for a number of experimental reasons. In other domains, such as social networks, only positive examples can be collected (such as 'liking' a particular product) because, by design, the negative labeling is not allowed. The development of classification models in this setting is often referred to as positive-unlabeled learning (Denis et al. 2005).

State-of-the-art techniques in positive-unlabeled learning tackle this problem by treating the unlabeled sample as negatives and training a classifier to distinguish between labeled (positive) and unlabeled examples. Following Elkan and Noto (2008), we refer to the classifiers trained on a labeled sample from the true distribution of inputs, containing both positive and negative examples, as *traditional classifiers*. Similarly, we refer to the classifiers trained on the labeled versus unlabeled data as *non-traditional classifiers*. In theory, the true performance of both traditional and non-traditional classifiers can be evaluated on a labeled sample from the true distribution (traditional evaluation). However, this is infeasible for non-traditional learners because such a sample is not available in positive-unlabeled learning. As a result, the non-traditional classifiers are evaluated by using the unlabeled sample as substitute for labeled negatives (non-traditional evaluation). Surprisingly, for a variety of performance criteria, non-traditional classifiers achieve similar performance under traditional evaluation as optimal traditional classifiers (Blanchard et al. 2010; Menon et al. 2015). The intuition for these results comes from the fact that in many practical situations, the posterior distributions in traditional and non-traditional setting provide the same optimal ranking of data points on a given test sample (Jain et al. 2016; Jain, White, and Radivojac 2016). Furthermore, the widely-accepted evaluation approaches using ROC or pr-rc curves are insensitive to the variation of raw prediction scores unless they affect the ranking.

Though the efficacy of non-traditional classifiers has been thoroughly studied (Peng et al. 2003; Elkan and Noto 2008; Ward et al. 2009; Menon et al. 2015), estimating their true performance has been much less explored. Such performance estimation often involves computing the fraction(s) of correctly and incorrectly classified examples from both classes; however, in absence of labeled negatives, the fractions computed under the non-traditional evaluation are incorrect, resulting in biased estimates. Figure 1 illustrates the effect of this bias by showing the traditional and non-traditional ROC curves on a handmade data set. Because some of the unlabeled examples in the training set are in fact positive, the area under the ROC curve estimated when the unlabeled examples were considered negative (non-traditional setting) underestimates the true performance for positive versus negative classification (traditional setting).

This paper formalizes and evaluates performance estimation of a non-traditional classifier in the traditional setting when the only available training data are (possibly noisy) positive examples and unlabeled data. We show that the true (traditional) performance of such a classifier can be recovered with the knowledge of class priors and the fraction of mislabeled examples in the positive set. We derive formulas for converting the ROC and pr-rc curves from the non-traditional to the traditional setting. Using these recovery formulas, we present methods to estimate true classification performance. Our experiments provide evidence that the methods for the recovery of a classifier's performance are sound and effective.

Problem formulation

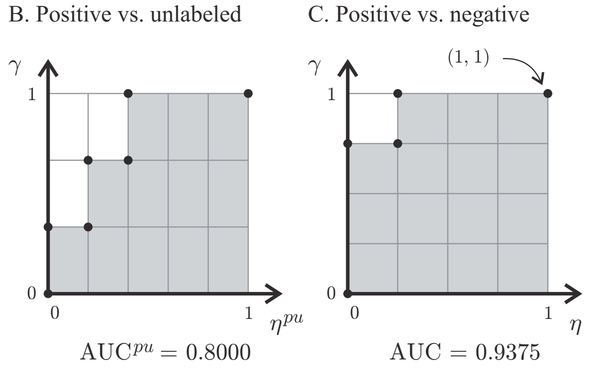
Consider a binary classification problem from input $x \in \mathcal{X}$ to output $y \in \mathcal{Y} = \{0, 1\}$ in a positive-unlabeled setting. Let f be the true distribution over the input space $\mathcal{X}$ from which the unlabeled sample is drawn and let f_1 and f_0 be the distributions of the positive and negative examples, respectively. It follows that f can be expressed as a two-component mixture containing f_1 and f_0 as

$$f(x) = \alpha f_1(x) + (1 - \alpha) f_0(x),$$

for all $x \in \mathcal{X}$ where $\alpha \in [0, 1)$ is the mixing proportion (positive class prior) giving the proportion of positives in f.

A. Data set: prediction scores and class labels

Prediction	Labeled	True class label	
0.986	yes, as 1	1	
0.943	no	1	*
0.863	yes, as 1	1	
0.789	no	0	
0.699	yes, as 1	1	
0.473	no	0	
0.211	no	0	
0.009	no	0	


B. Positive vs. unlabeled

C. Positive vs. negative

Figure 1: Illustration of the difference in classifier evaluation. (A) A data set with eight examples, three labeled positive and five unlabeled. One unlabeled example is positive (marked by an asterisk outside the table), whereas four are negative. A prediction score between zero and one is provided for each example. (B) The ROC plot (γ = true positive rate; η = false positive rate) when all unlabeled examples are considered negative. (C) The true ROC plot where all examples are correctly labeled. The areas under the ROC curves are calculated without interpolation as the total area of shaded boxes.

Let now g be the distribution over $\mathcal{X}$ from which the labeled sample is drawn. We similarly express g as a two-component mixture containing f_1 and f_0 as

$$g(x) = \beta f_1(x) + (1 - \beta) f_0(x),$$

for all $x \in \mathcal{X}$ where $\beta \in (\alpha, 1]$ gives the proportion of positives in labeled data. All labeled examples are labeled as positives; thus, when $\beta = 1$ we say that the labeled data is clean. When $\beta < 1$, the labeled data contains a fraction $(1 - \beta)$ of negatives that are in this case mislabeled. We will refer to the latter scenario as the noisy positive setting.

Let X_1 be the (positively) labeled sample drawn according to $g(x)$ and X be the unlabeled sample drawn according to $f(x)$. The learning objective is to train a classifier that discriminates between positive and negative data and estimate its performance. However, we can only train a non-traditional classifier $h : \mathcal{X} \to \mathcal{Y}$ between labeled and unlabeled data and estimate its performance by considering that all labeled data are positive and all unlabeled data are negative. We refer to the performance of $h(x)$ directly estimated from samples X_1 and X as perf^{pu}. Given a non-traditional classifier $h(x)$ and its performance perf^{pu}, the main goal of this work is to estimate (recover) its performance in the traditional setting; i.e., its performance as a discriminator between positive and negative data.

Methods

We consider a family of binary classifiers that map $\mathcal{X}$ into $\mathcal{Y}$. To simplify the presentation, we can think of the entire family as generated from a single model that maps $\mathcal{X}$ into $\mathbb{R}$, where each individual classifier corresponds to a decision threshold picked from $\mathbb{R}$. The classifier gives the positive class '1' when the model's output is above the threshold and the negative class '0' otherwise.

The true positive rate (sensitivity, recall) of each classifier is defined as the probability of correctly predicting a positive example; the true negative rate (specificity) is defined as the probability of correctly predicting a negative example; the false positive rate is defined as $1 -$ specificity, and the false negative rate is defined as $1 -$ sensitivity. Finally, the precision is defined as the probability that a positive prediction is correct; conversely, the false discovery rate is defined as $1 -$ precision (Hastie et al. 2001). Given a test set, each of the quantities above is estimated using relative frequencies. In this setup, each classifier corresponds to a single confusion matrix, whereas the entire family of classifiers corresponds to a particular ROC curve and a particular pr-rc curve (Fawcett 2006). The two main performance criteria considered in this work are the area under the ROC curve and the area under the pr-rc curve.

The case of clean positive data

We first consider the setting of clean positive data, where the labeled data does not incorrectly contain negatives ($\beta = 1$), to provide intuition before moving to the more general noisy-positive setting. For a classifier $h : \mathcal{X} \to \mathcal{Y}$, the true positive rate, γ, and false positive rate, η, can be defined as

$$\begin{aligned} \gamma &= \mathbb{E}_{f_1}[h(x)] \\ \eta &= \mathbb{E}_{f_0}[h(x)], \end{aligned}$$

where $\mathbb{E}_f$ denotes expectation with respect to a distribution f. The goal is to estimate these values, despite the fact that we only have access to positive labels.

The true positive rate can be estimated as the empirical mean of $h(x)$ over the positively labeled sample X_1

$$\hat{\gamma} = \frac{1}{|X_1|} \sum_{x \in X_1} h(x)$$

because X_1 was sampled from f_1. The false positive rate, however, cannot be so simply estimated, because we do not have access to a sample from f_0. Further, this prevents the estimation of the ROC curve and the area under this curve (AUC). Typically, ROC curves and AUCs are reported based only on the performance of the non-traditional positive-unlabeled classifier, h, on discriminating between positives and unlabeled data. The ROC curve for the positive versus unlabeled classification, ROC^{pu}, can be estimated by plotting $\hat{\gamma}$ against $\hat{\eta}^{pu}$ across different classifiers, where $\hat{\eta}^{pu}$, an estimate of $\eta^{pu} = \mathbb{E}_f[h(x)]$, can be estimated using the unlabeled sample (which corresponds to the negative sample for the non-traditional positive-unlabeled classifier h):

$$\hat{\eta}^{pu} = \frac{1}{|X|} \sum_{x \in X} h(x).$$

This curve, however, does not represent the true performance of h for positive versus negative classification. Similar difficulties exist in estimating the precision

$$\rho = \frac{\alpha \mathbb{E}_{f_1}[h(x)]}{\mathbb{E}_f[h(x)]}$$

that requires the positive class prior α, though recall, which is equal to γ, can be directly estimated.

Of key interest, therefore, is a correction approach that provides an estimate of the true performance. We provide just such a result in Theorem 1 below for the more general setting of noisy positives (see next Section). Using this theorem for $\beta = 1$, for example, we can express the false positive rate η in terms of the positive-unlabeled false positive rate, η^{pu} as[1]

$$\eta = \frac{\eta^{pu} - \alpha\gamma}{1 - \alpha},$$

and the AUC of the classifier on the positive-negative classification problem in terms of the AUC of the classifier on the positive-unlabeled classification problem[2]:

$$\text{AUC} = \frac{\text{AUC}^{pu} - \frac{\alpha}{2}}{1 - \alpha}.$$

Therefore, given estimates of $\alpha, \gamma, \eta^{pu}$ and AUC^{pu}, we can obtain estimates of AUC and the precision. In the next Section, we present this key result that enables this conversion and also shows that the estimated AUC is better than AUC^{pu}.

The case of (possibly) noisy positive data

In this section we consider a more general case where the labeled sample of positives is allowed to be noisy; i.e., some positives may actually be negatives. Since this setting is a strict generalization of the previous discussion, we will overload terminology and use η^{pu} again as the positive-unlabeled false positive rate.

In addition to previous difficulties, we now also cannot estimate the true positive rate γ, because we do not have access to an unbiased sample from f_1; rather, we only have access to a sample contaminated with negatives. Nonetheless, we can express all of the desired rates in terms of only rates for the non-traditional classifier.

Theorem 1. *For a given classifier $h : \mathcal{X} \to \mathcal{Y}$, the true positive rate γ and the false positive rate η can be expressed*

[1] Iakoucheva et al. (2004) also provide this result for uncorrupted positive data.

[2] Menon et al. (2015) provide an equivalent formula for the AUC. In Theorem 1, we give a full derivation from the probabilistic definition of the AUC and conversion formulas for other measures.

in terms of the positive-unlabeled γ^{pu} *and* η^{pu}

$$\gamma = \frac{(1-\alpha)\gamma^{pu} - (1-\beta)\eta^{pu}}{\beta - \alpha} \tag{1}$$

$$\eta = \frac{\beta\eta^{pu} - \alpha\gamma^{pu}}{\beta - \alpha}. \tag{2}$$

The precision ρ *can either be converted from a positive-unlabeled precision* ρ^{pu}*, with* $c = {}^{|X_1|}/_{(|X|+|X_1|)}$*, as*

$$\rho = \frac{\alpha(1-\alpha)}{\beta-\alpha}\left(\frac{1-c}{c}\left(\frac{\rho^{pu}}{1-\rho^{pu}}\right) - \frac{1-\beta}{1-\alpha}\right)$$

or computed directly as

$$\rho = \frac{\alpha\gamma}{\eta^{pu}}. \tag{3}$$

Further, consider a family of classifiers $\mathcal{F} = \{h_\eta\}$ *indexed by* $\eta \in [0,1]$ *where* η *is the false positive rate of* h_η*. Then for the ROC curve obtained from varying* η*, the AUC can be expressed in terms of the positive-unlabeled AUC*^{pu} *as*

$$AUC = \frac{AUC^{pu} - \frac{1-(\beta-\alpha)}{2}}{\beta - \alpha}. \tag{4}$$

Moreover, AUC $>$ *AUC*^{pu}*, if and only if AUC*$^{pu} > {}^1/_2$ *and* $\beta - \alpha < 1$.

Proof.

$$\begin{aligned}\eta^{pu} &= \mathbb{E}_f[h(x)] \\ &= \alpha\mathbb{E}_{f_1}[h(x)] + (1-\alpha)\mathbb{E}_{f_0}[h(x)] \\ &= \alpha\gamma + (1-\alpha)\eta.\end{aligned}$$

Similarly, we can obtain the true positive rate

$$\begin{aligned}\gamma^{pu} &= \mathbb{E}_g[h(x)] \\ &= \beta\mathbb{E}_{f_1}[h(x)] + (1-\beta)\mathbb{E}_{f_0}[h(x)] \\ &= \beta\gamma + (1-\beta)\eta.\end{aligned}$$

We can then solve for η and γ to get the result.

Next, we consider the precision. We can directly re-express the precision as

$$\rho = \frac{\alpha\mathbb{E}_{f_1}[h(x)]}{\mathbb{E}_f[h(x)]} = \frac{\alpha\gamma}{\eta^{pu}}.$$

To obtain a conversion from ρ^{pu}, first consider

$$\begin{aligned}\rho^{pu} &= \frac{c\mathbb{E}_g[h(x)]}{\mathbb{E}_{cg+(1-c)f}[h(x)]} \\ &= \frac{c\mathbb{E}_g[h(x)]}{c\mathbb{E}_g[h(x)] + (1-c)\mathbb{E}_f[h(x)]} \\ &= \frac{1}{1 + \frac{1-c}{c}\frac{\mathbb{E}_f[h(x)]}{\mathbb{E}_g[h(x)]}}\end{aligned}$$

We can express a component of this as

$$\begin{aligned}\frac{\mathbb{E}_g[h(x)]}{\mathbb{E}_f[h(x)]} &= \frac{\beta\mathbb{E}_{f_1}[h(x)] + (1-\beta)\mathbb{E}_{f_0}[h(x)]}{\mathbb{E}_f[h(x)]} \\ &= \frac{\beta}{\alpha}\frac{\alpha\mathbb{E}_{f_1}[h(x)]}{\mathbb{E}_f[h(x)]} + \frac{1-\beta}{1-\alpha}\frac{(1-\alpha)\mathbb{E}_{f_0}[h(x)]}{\mathbb{E}_f[h(x)]} \\ &= \frac{\beta}{\alpha}\rho + \frac{1-\beta}{1-\alpha}(1-\rho) \\ &= \frac{\beta-\alpha}{\alpha(1-\alpha)}\rho + \frac{1-\beta}{1-\alpha}\end{aligned}$$

where rearranging gives the result.

Next, we derive an equation that allows estimation of the AUC directly from the AUC^{pu}, α and β. Consider a family of classifiers $\mathcal{F} = \{h_\eta\}$ indexed by $\eta \in [0,1]$ where η is the false positive rate of h_η. We can express the $\gamma, \eta^{pu}, \gamma^{pu}$ of h_η as a function of η as follows:

$$\begin{aligned}\gamma(\eta) &= \mathbb{E}_{f_1}[h_\eta(x)], \\ \eta^{pu}(\eta) &= \mathbb{E}_f[h_\eta(x)] \\ &= \alpha\gamma(\eta) + (1-\alpha)\eta, \\ \gamma^{pu}(\eta) &= \mathbb{E}_g[h_\eta(x)] \\ &= \beta\gamma(\eta) + (1-\beta)\eta.\end{aligned}$$

By definition, the expression for AUC^{pu} is

$$\begin{aligned}\text{AUC}^{pu} &= \int_0^1 \gamma^{pu}(\eta)\frac{d\eta^{pu}(\eta)}{d\eta}d\eta \\ &= \int_0^1 (\beta\gamma(\eta) + (1-\beta)\eta)\left(\alpha\frac{d\gamma(\eta)}{d\eta} + (1-\alpha)\right)d\eta \\ &= \alpha\beta\int_0^1 \gamma(\eta)\frac{d\gamma(\eta)}{d\eta}d\eta + (1-\alpha)\beta\int_0^1 \gamma(\eta)d\eta \\ &\quad + \alpha(1-\beta)\int_0^1 \eta\frac{d\gamma(\eta)}{d\eta}d\eta + (1-\alpha)(1-\beta)\int_0^1 \eta d\eta\end{aligned}$$

Now solving for each integral, we obtain

$$\begin{aligned}\text{AUC}^{pu} &= \frac{\alpha\beta}{2}[\gamma^2(1) - \gamma^2(0)] + (1-\alpha)\beta\text{AUC} \\ &\quad + \alpha(1-\beta)\left[[\eta\gamma(\eta)]_0^1 - \int_0^1 \gamma(\eta)d\eta\right] \\ &\quad + \frac{(1-\alpha)(1-\beta)}{2}[1^2 - 0^2] \\ &= \frac{\alpha\beta + 2\alpha(1-\beta) + (1-\alpha)(1-\beta)}{2} \\ &\quad + [(1-\alpha)\beta - \alpha(1-\beta)]\text{AUC} \\ &= \frac{1-(\beta-\alpha)}{2} + (\beta-\alpha)\text{AUC}\end{aligned}$$

Rearranging the terms gives the desired result. Finally, from Equation 4, we see that

$$\text{AUC} - \text{AUC}^{pu} = \frac{1-(\beta-\alpha)}{\beta-\alpha}\left(\text{AUC}^{pu} - \frac{1}{2}\right)$$

proving AUC $>$ AUC^{pu}, if and only if AUC$^{pu} > {}^1/_2$ and $\beta - \alpha < 1$. □

Experiments and results

Data sets and classification models

Our estimators were evaluated using twelve real-life data sets from the UCI Machine Learning Repository (Lichman 2013). All data sets were appropriately modified for binary classification; e.g., regression problems were converted into classification problems based on the mean of the target variable, whereas multiclass classification problems were converted into binary problems by combining classes. When needed, categorical features were converted into numerical features based on the sparse binary representation.

Classifiers were constructed as ensembles of 100 feed-forward neural networks (Breiman 1996). Each network had five hidden neurons and was trained using resilient propagation (Riedmiller and Braun 1993). A validation set containing 25% of the training data was used to terminate training. For simplicity, no training parameters were varied. Accuracies were estimated using the out-of-bag approach.

Experimental protocols

To evaluate the quality of performance estimation we first established the ground truth performance of a model by estimating accuracy in a standard supervised setting. All positive examples in all data sets were considered positive and all negative examples were considered negative. A model was then constructed and evaluated for its performance.

We next simulated the positive-unlabeled setting where we randomly included 1,000 examples (or 100 for smaller data sets) in the positive data set X_1. The number of actual positive examples in each labeled set was a function of parameter $\beta \in \{1, 0.95, 0.75\}$. For example, when $\beta = 1$, all positively labeled examples were positive, and when $\beta < 1$, an appropriate fraction of the (positively) labeled data set X_1 was filled with negatives. The remaining examples (positive and negative) were declared unlabeled (data set X). The size of the unlabeled data was limited to 10,000 (where relevant) and the fraction of positives in the unlabeled data was used as true α. Using all positively labeled examples as positives and all unlabeled examples as negatives, we then estimated the performance of the model in the positive-unlabeled setting. All experiments were repeated fifty times by randomly selecting positives and negatives for the labeled data.

We used our methodology from the previous Section to recover the true accuracy of a model. To recover the area under the ROC curve, we used the *direct conversion* (D) from Equation 4 as well as *indirect conversion* (I) where traditional true positive and false positive rates were recovered using Equations 1-2 for every threshold and then used to reconstruct the ROC curve. In the case of recovering the pr-rc curve, only the indirect conversion was used (using Equations 1 and 3) as no direct conversion formula is known to us. The full algorithm for the indirect recovery is given in the arXiv supplement of this paper.

All experiments were carried out (i) by assuming that the class prior α and noise fraction β were known (R), and (ii) by estimating α and β from positive and unlabeled data (E). These experiments were carried out to quantify the performance loss due to the inability to perfectly estimate (α, β). Class priors and noise fraction were estimated using the AlphaMax algorithm (Jain et al. 2016; Jain, White, and Radivojac 2016). Several recent studies have determined good performance of AlphaMax (Jain et al. 2016; Jain, White, and Radivojac 2016; Ramaswamy et al. 2016), in both clean and noisy setting.

The direct recovery methods using real and estimated (α, β) are hereafter referred to as DR and DE methods, respectively, whereas the indirect recovery methods are similarly referred to as IR and IE methods. All four approaches were used to evaluate the estimated AUCs and only IR and IE methods were used to evaluate the estimated area under the pr-rc curve (AUC-PR).

Results

Figure 2 shows the general trends in estimating AUC and AUC-PR over all data sets. Detailed dataset-specific evaluations over all summary statistics are given in Tables 1-2, while the error between the true and recovered performance is further characterized in Figures 3-4. Tables 1-2 and Figures 3-4 are shown in the arXiv supplement of this paper.

Figure 2(a) shows that, as expected, AUC^{pu} consistently underestimates the true performance. Moreover, it deteriorates with increase in noise. On the other hand, using the correct values for α and β (IR and DR, corresponding to the yellow and green boxes) leads to excellent performance over all values of β. Replacing the true (α, β) by their estimates obtained from AlphaMax did not lead to significantly different performance estimates (IE and DE, corresponding to the blue and purple boxes). Since class prior estimation guarantees identifiability of only the upper bounds of (α, β), the observed differences are reasonable. Although the aggregate performance of direct and indirect estimation is similar, a detailed comparison between these methods (DR vs. IR and DE vs. IE) provides evidence that the indirect method was superior in both cases ($P = 6.5 \cdot 10^{-6}$ for real α and β and $P = 5.7 \cdot 10^{-3}$ for estimated α and β; one-sided binomial test). Full details of these comparisons are shown in the arXiv supplement.

Figure 2(b) shows that the performance breaks down with increase in the absolute error of estimates of $\beta - \alpha$. We selected this criterion because the term $\beta - \alpha$ appears in the denominator of Equation 4 and thus could significantly influence the quality of performance. The increase in error more strongly affects the estimators with approximate (α, β). Interestingly, the estimators IR and DR both underestimate, and IE and DE both overestimate. We note that in some cases the data sets obtained from UCI Machine Learning Repository may not be perfectly labeled in the first place. This suggests that our ground truth performance might be slightly biased for some data sets which would lead to a situation that the estimated performance is in fact more accurate than observed.

Figures 2(c) and 2(d) show the equivalent plots for AUC-PR from which similar conclusions can be drawn. However, errors in the uncorrected AUC-PR estimates (red boxes) are much higher in comparison. Estimating AUC-PR is therefore not particularly meaningful in the non-traditional setting because precision is sensitive to the proportion of labeled positives in the data set; i.e., $|X_1|/(|X|+|X_1|)$, whereas γ and η are not.

Related work

Evaluation metrics

Two-dimensional performance characterization such as ROC or pr-rc curves and the summary statistics based on them have become mainstream in empirical evaluation of classification performance (Flach 2003; Fawcett 2006;

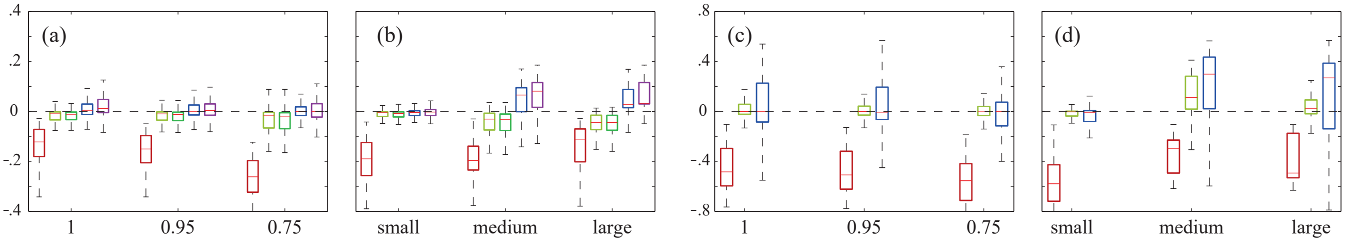


Figure 2: The distribution of error of AUC (a, b) and AUC-PR (c, d) estimators on the data generated from the 12 datasets. PU represents the estimates on the positive unlabeled data without correction. IR, DR, IE, DE are the corrected estimates, either using the **R**eal values of (α, β) or the **E**stimated values. **D** indicates that the AUC^{pu} was **D**irectly corrected using equation (4) (direct conversion is not done for AUC-PR) and **I** indicates **I**ndirect correction by first correcting for the ROC or pr-rc curves. AUC estimates above 1 were clipped. The x-axis for the left column is the real value of β (in increasing noise order) and for the right column it is the absolute error of $\beta - \alpha$ estimate binned into small: $[0, 0.1)$, medium: $[0.1, 0.2)$ and large: $[0.2, \infty)$.

Davis and Goadrich 2006; Boyd et al. 2012; Clark and Radivojac 2013; Flach and Kull 2015). Of particular interest to our work is the well-explored relationship between these performance metrics and class priors. For example, Hernández-Orallo et al. (2012) use class priors and area under the ROC curve to compute the expected classification accuracy, whereas Boyd et al. (2012) relate class priors to the size of the unachievable region in pr-rc space. In the domain of positive-unlabeled learning, Menon et al. (2015) give the relationship between traditionally and non-traditionally evaluated balanced error rates and AUCs of a given classifier. They use this relationship to demonstrate that constructing a non-traditional classifier by optimizing non-traditional AUC results in an optimal traditional AUC. Claesen et al. (2015) similarly argue the importance of class priors and show how to compute bounds on the true ROC or pr-rc curves. In contrast, our approach directly estimates the unknown statistics and derives a closed-form conversion formula for recovering the area under the ROC curve from the first principles. Another similar work, although in the area of structured-output learning, is by Jiang et al. (2014) who studied the impact of sequential completion of the (structured) target variable; however, their work makes fewer assumptions on the data distributions and does not lead to the recovery of true performance.

Class prior and noise estimation

Though class prior (α) estimation in positive-unlabeled learning is nontrivial, several algorithms have recently emerged in the literature. Elkan and Noto (2008) estimate the priors from the probability obtained by calibrating the scores of a non-traditional classifier under strong assumptions that the class-conditional distributions do not overlap. The same assumptions are used by (du Plessis and Sugiyama 2014) who estimate the class prior as the minimizer of the Pearson divergence. du Plessis et al. (2015) improve the method by using penalized f-divergence to allow overlap. Blanchard et al. (2010) and Jain et al. (2016) showed that class prior estimation, in general, is an ill-posed problem and introduce an "irreducibility" constraint on the distribution of the negatives that makes the problem well defined. Blanchard et al. (2010) estimate the class prior as the slope of the right endpoint of the empirical ROC curve from non-traditional classifiers while Sanderson and Scott (2014) use a fitted curve instead of the actual ROC curve to smooth large noise at endpoints. Loosely speaking, the ROC approach is based on the fact that the class prior under the irreducibility assumption is the minimum value attained by the ratio of the unlabeled and positive sample densities (Jain et al. 2016). Jain et al. (2016) also give an algorithm, AlphaMax, a nonparametric maximum likelihood based approach suitable for high-dimensional data. Ramaswamy et al. (2016) give an algorithm based on embedding distributions into a reproducing kernel Hilbert spaces.

In the case of noisy positives, Scott et al. (2013) and Jain, White, and Radivojac (2016) impose a "mutual irreducibility" constraint on the distribution of positives and negatives, to make the class prior and the noise proportion estimation well defined. Jain, White, and Radivojac (2016) estimate α, β by combining the outputs of two executions of AlphaMax, one of which flips the role of positive and unlabeled samples.

Conclusions

In this paper we propose simple methods for correcting the estimated performance of classifiers trained in the positive-unlabeled setting. We prove a fundamental result about the relationship between widely-used performance measures and their positive-unlabeled counterparts. The resulting estimators were evaluated over a diverse group of data sets to show that it is feasible and practical to obtain accurate estimates of a classifier's performance in the task of discriminating positive and negative examples.

The corrected performance measures were uniformly more accurate than the positive-unlabeled estimates, which typically underestimated the performance. Furthermore, we showed that the indirect method for performance recovery outperformed the direct method. This notwithstanding, we do not recommend stopping the established practice of reporting perf^{pu}; rather we propose that the corrected performance measures should also be provided. In domains where α and β are unknown, such estimates will contribute to a better understanding of a classifier's performance and a deeper understanding of the domain itself.

Acknowledgements

We thank Prof. Michael W. Trosset, Kymerleigh A. Pagel and Vikas Pejaver for helpful comments. Grant support: NSF DBI-1458477, NIH R01MH105524, NIH R01GM103725, and the Indiana University Precision Health Initiative.

References

Blanchard, G.; Lee, G.; and Scott, C. 2010. Semi-supervised novelty detection. *J Mach Learn Res* 11:2973–3009.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, COLT 1998, 92–100.

Boyd, K.; Costa, V. S.; Davis, J.; and Page, C. D. 2012. Unachievable region in precision-recall space and its effect on empirical evaluation. In *Proceedings of the 29th International Conference on Machine Learning*, ICML 2012, 639–646.

Breiman, L. 1996. Bagging predictors. *Mach Learn* 24:123–140.

Claesen, M.; Davis, J.; De Smet, F.; and De Moor, B. 2015. Assessing binary classifiers using only positive and unlabeled data. *arXiv preprint arXiv:1504.06837*.

Clark, W. T., and Radivojac, P. 2013. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* 29(13):i53–i61.

Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT 2008, 38–53.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML 2006, 233–240.

Denis, F.; Gilleron, R.; and Letouzey, F. 2005. Learning from positive and unlabeled examples. *Theor Comput Sci* 348(16):70–83.

du Plessis, M. C., and Sugiyama, M. 2014. Class prior estimation from positive and unlabeled data. *IEICE Trans Inf & Syst* E97-D(5):1358–1362.

du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2015. Class-prior estimation for learning from positive and unlabeled data. In *Proceedings of the 7th Asian Conference on Machine Learning*, volume 45 of *ACML 2015*, 221–236.

Elkan, C., and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2008, 213–220.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, IJCAI 2001, 973–978.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recogn Lett* 27:861–874.

Flach, P. A., and Kull, M. 2015. Precision-recall-gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems*, NIPS 2015, 838–846.

Flach, P. A. 2003. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *Proceedings of the 20th International Conference on Machine Learning*, ICML 2003, 194–201.

Hanley, J., and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36.

Hastie, T.; Tibshirani, R.; and Friedman, J. H. 2001. *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer Verlag.

Hernández-Orallo, J.; Flach, P.; and Ferri, C. 2012. A unified view of performance metrics: translating threshold choice into expected classification loss. *J Mach Learn Res* 13(1):2813–2869.

Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; and Dunker, A. K. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32(3):1037–1049.

Jain, S.; White, M.; Trosset, M. W.; and Radivojac, P. 2016. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*.

Jain, S.; White, M.; and Radivojac, P. 2016. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, NIPS 2016, 2685–2693.

Jiang, Y.; Clark, W. T.; Friedberg, I.; and Radivojac, P. 2014. The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. *Bioinformatics* 30(17):i609–i616.

Lichman, M. 2013. UCI Machine Learning Repository.

Menon, A. K.; van Rooyen, B.; Ong, C. S.; and Williamson, R. C. 2015. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML 2015, 125–134.

Peng, K.; Vucetic, S.; Han, B.; Xie, H.; and Obradovic, Z. 2003. Exploiting unlabeled data for improving accuracy of predictive data mining. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, ICDM 2003, 267–274.

Ramaswamy, H. G.; Scott, C.; and Tewari, A. 2016. Mixture proportion estimation via kernel embedding of distributions. *arXiv preprint arXiv:1603.02501*.

Riedmiller, M., and Braun, H. 1993. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*, ICNN 1993, 586–591.

Saerens, M.; Latinne, P.; and Decaestecker, C. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Comput* 14:21–41.

Sanderson, T., and Scott, C. 2014. Class proportion estimation with application to multiclass anomaly rejection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, AISTATS 2014, 850–858.

Scott, C.; Blanchard, G.; and Handy, G. 2013. Classification with asymmetric label noise: consistency and maximal denoising. *J Mach Learn Res W&CP* 30:489–511.

Ward, G.; Hastie, T.; Barry, S.; Elith, J.; and Leathwick, J. 2009. Presence-only data and the EM algorithm. *Biometrics* 65(2):554–563.

Whalen, A. D. 1971. *Detection of signals in noise*. New York, NY: Academic Press.