Pattern Recognition Letters 85 (2017) 35–41

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

ROC curves and nonrandom data

Jonathan Aaron Cook[1]

Public Company Accounting Oversight Board, 1666 K Street, NW, Washington, DC 2006, USA

ARTICLE INFO

Article history:
Received 23 May 2016
Available online 25 November 2016

MSC:
41A05
41A10
65D05
65D17

Keywords:
ROC curves
Classifier evaluation
Sample-selection bias

ABSTRACT

This paper shows that when a classifier is evaluated with nonrandom test data, ROC curves differ from the ROC curves that would be obtained with a random sample. To address this bias, this paper introduces a procedure for plotting ROC curves that are inferred from nonrandom test data. I provide simulations to illustrate the procedure as well as the magnitude of bias that is found in empirical ROC curves constructed with nonrandom test data. The paper also includes a demonstration of the procedure on (non-simulated) data used to model wine preferences in the wine industry.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In many settings, data are collected in a nonrandom fashion. The decision to investigate insurance claims for fraud may be based on a predictive model. Investigating insurance claims is costly and it may be difficult to allocate resources to inspect a random sample of claims. Similarly, the Internal Revenue Service (IRS) uses a model that predicts tax-filing errors to select tax returns for audits. A recommender system may only show the user items that are predicted to be of interest. In these three examples, data are only collected for instances that are judged to be more likely to be positive cases.

This paper makes two contributions. This paper's first contribution is a characterization of the bias that results in receiver operating characteristic (ROC) curves when they are constructed with nonrandom test data.[2] The bias described by this paper is caused by constructing the ROC curve with test data that are not representative of the population of interest. This paper does not consider the effects of using test data that are not representative of the training data. There is a downward bias for ROC curves when the classifier is strongly correlated with the classifier that was used to select the test data. By contrast, ROC curves are pushed outward for a classifier with low correlation to the classifier that was used to select the test data. The bias that arises from using another classifier to select the test data is related to (but different from) sample-selection bias for linear regression, which has been studied in the econometric literature.

This paper's second contribution is a procedure to create ROC curves that provide a consistent estimate of the ROC curve that would be obtained with random test data. This procedure infers the predictive power of the classifier based on available data and plots the implied ROC curve. The inferred ROC curves are based on econometric work on bivariate probit analysis (e.g. [21] and [19]). A key difference between this paper and prior work on selection problems is that the problems considered by this paper are not regression equations. Section 5 discusses instances for which ROC curves are biased, but the parameters of a regression equation would not be.

I make distributional assumptions that lead to a maximum likelihood problem that is similar to those encountered in estimating regression equations with sample selection. A classifier's expected ROC curve is determined by two parameters. The first parameter determines how many positive cases there are in the population. The second parameter is the correlation of the classifier's output for each instance with that instance's latent propensity to be a positive case.

The presented procedure is related to the Dorfman–Alf [6] procedure for estimating parameters of fitted ROC curves, which also uses maximum likelihood estimates under parametric assumptions. (Extensions of the Dorfman–Alf procedure include [17], [24],

E-mail address: jacook@uci.edu

[1] The PCAOB, as a matter of policy disclaims responsibility for any private publication or statement by any of its Economic Research Fellows and employees. The views expressed in this paper are the views of the author and do not necessarily reflect the views of the Board, individual Board members, or staff of the PCAOB.

[2] Throughout this paper, I refer to data that are used to evaluate a classifier's performance as "test data." Data that are used to train the classifier are referred to as "training data."

http://dx.doi.org/10.1016/j.patrec.2016.11.015
0167-8655/© 2016 Elsevier B.V. All rights reserved.

Table 1
Confusion matrix.

		Truth	
		Positive	Negative
Prediction	Positive	True Positives (*TP*)	False Positives (*FP*)
	Negative	False Negatives (*FN*)	True Negatives (*TN*)
	Total	Positives (*P*)	Negatives (*N*)

and [7].) The Dorfman–Alf procedure and its various extensions do not correct for selection bias.

This paper contributes to the literature on evaluating classifiers. Recent works have shown the connections between ROC curves and precision-recall curves [5] and cost curves [13]. Other work on the properties of evaluation metrics for classifiers includes Wang et al. [22], who show that normalized discounted cumulative gains (NDCG) can consistently distinguish classifiers, and Moffat [18], who provides properties of evaluation metrics. There does not appear to be any existing work on evaluating classifiers with nonrandom data.

Training a classifier with nonrandom data is beyond the scope of this paper. This paper does not discuss the effects of having nonrandom training data. To create classifiers with nonrandom training data, the econometric literature has built on the sample-selection correction regression of Heckman [11,12] (see [21] for a binary classifier). The credit-scoring literature has introduced *reject inference*, which incorporates information from unselected items, to improve classifier performance (see, for example, [4]).

In the next section, I introduce notation and derive the bias in ROC curves when the classifier being evaluated was used to select the test data. I derive a ROC curve that consistently estimates the ROC curve that would be obtained with random data in Section 3. Sections 4 and 5 present an example and Monte Carlo simulations to illustrate this procedure as well as the bias found in empirical ROC curves. Section 6 concludes.

2. Classifiers and ROC curves

A *classifier* maps instances to predicted classes. This paper focuses on *binary classifiers*, which map to two classes (e.g., positive and negative). While some classifiers map directly to predicted classes, this paper focuses on classifiers that produce a continuous output. Given the classifier's output and a threshold, we classify all instances above the threshold as positive and all instances below the threshold as negative.

The confusion matrix in Table 1 defines true positives (TP), true negatives (TN), positives (P), and negatives (N). Sensitivity and specificity are defined as

$$\text{Sensitivity} = \frac{TP}{P}, \quad \text{and} \tag{1}$$

$$\text{Specificity} = \frac{TN}{N}. \tag{2}$$

ROC curves, which plot sensitivity as a function of specificity for all possible thresholds,[3] illustrate a classifier's trade-off between true positives and false negatives. A higher value of sensitivity for a given value of specificity indicates better performance. The area under the ROC curve (AUC) is a commonly used metric for evaluating a classifier's performance (as described by Bradley [1]). If the classifier's output has no connection to the true class, the expected AUC would be .5. An excellent introduction to ROC curves is provided by Fawcett [8].

[3] The thresholds are often referred to as "operating points."

Evaluating a classifier with nonrandom test data

This section introduces notation and provides some analytical results regarding the sample-selection bias for ROC curves. Let us denote the continuous output of classifier $\mathcal{A}$ for each instance i as a_i. I assume that there is some unobserved propensity to be a positive case and denote this propensity as p_i for each instance i. The true classification of each instance is

$$\text{outcome}_i = \begin{cases} \text{positive} & \text{if } p_i \geq p^* \\ \text{negative} & \text{otherwise} \end{cases}, \tag{3}$$

where p^* is the threshold for an instance to be a positive case. A value of $p^* = 0$ indicates that half of the observations are positive cases. The class skew increases with the absolute value of p^*. Throughout this paper, I treat both p_i and a_i as (possibly correlated) random variables. The modeler never observes p_i, only outcome_i. For a given threshold c, we can give probabilistic definitions of sensitivity and specificity:

$$\text{Sensitivity} = \text{Prob}(a_i > c \mid p_i > p^*), \quad \text{and} \tag{4}$$

$$\text{Specificity} = \text{Prob}(a_i < c \mid p_i < p^*). \tag{5}$$

The values in Eqs. (1) and (2) provide sample estimates of these probabilities.

Another classifier, $\mathcal{B}$ with output denoted as b, is used to select the test data. This paper focuses on situations in which b is not observed. Appendix B explores the situation of an observed b. I assume that each instance of b can be written as

$$b_i = \delta X_i + \gamma a_i + \varepsilon_i,$$

where X_i is a vector of features for case i and ε_i is a standard normal random variable. The parameter δ is a vector of coefficients and γ indicates the degree to which the classifier's output was incorporated into the selection process. I assume that ε is mean independent of X and α, i.e. $E(\varepsilon|X, \alpha) = 0$. This assumption allows for estimation of δ and γ by a probit regression.

Data is selected according to the rule

$$\begin{cases} \text{Selected} & \text{if } \delta X_i + \gamma a_i + \varepsilon_i > s \\ \text{Not selected} & \text{otherwise} \end{cases}, \tag{6}$$

where s is a constant. Sensitivity and specificity conditional on selection are denoted as

$$\text{Sensitivity} \mid \text{Selection} = \text{Prob}(a_i > c \mid p_i > p^*, b_i > s) \tag{7}$$

$$\text{Specificity} \mid \text{Selection} = \text{Prob}(a_i < c \mid p_i < p^*, b_i > s). \tag{8}$$

When data are chosen based on a classifier's output, the estimates in Eqs. (1) and (2) provide an estimate of the values in Eqs. (7) and (8) instead of the values in Eqs. (4) and (5).

To build our intuition about the effect of nonrandom data, I briefly digress to consider a simpler form of choosing test data based on a classifier: selecting the test data using the classifier that we want to evaluate. Sensitivity and specificity conditional on selection on the classifier to be evaluated are denoted as

$$\text{Sensitivity} \mid \text{Selection} = \text{Prob}(a_i > c \mid p_i > p^*, a_i > s) \tag{9}$$

$$\text{Specificity} \mid \text{Selection} = \text{Prob}(a_i < c \mid p_i < p^*, a_i > s). \tag{10}$$

The following lemma will aid in proving our results regarding the bias in empirical ROC curves for test data that are selected by the classifier that we want to evaluate.

Lemma 1. *For a fixed value of c, conditioning on selection by the classifier that we want to evaluate*

(i) *Increases sensitivity, i.e.*

$$Sensitivity < Sensitivity \mid Selection \quad for\ all - \infty < s < c,$$

and

(ii) *Decreases specificity, i.e.*

$$Specificity > Specificity \mid Selection \quad for\ all - \infty < s < c.$$

All proofs are provided in Appendix A. For a given cutoff level, selection moves sensitivity and specificity in opposite directions. The intuition for this result is that, as we focus on instances that our classifier considers more likely to be positive cases, we will have more positive cases in our test data. Sensitivity, which is conditional on the number of positive cases, is biased downward as the relative prevalence of positive cases increases. Similarly, specificity is biased upward as the relative number of negative cases decreases.

The ROC curve plots sensitivity as a function of specificity:

$$\text{Sensitivity}(\text{Specificity}) = \text{Prob}(a_i > c \mid p_i > p^*),$$

where c satisfies

$$\text{Specificity} = \text{Prob}(a_i < c \mid p_i < p^*).$$

Up to this point, this paper has not made any distributional assumptions. To derive analytical results about the effect of selection on ROC curves, it is useful to assume that p_i and a_i come from a bivariate normal distribution:

$$\begin{pmatrix} p_i \\ a_i \end{pmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} \\ \rho_{ap} & 1 \end{bmatrix}\right).$$

The multivariate normal distribution is chosen because of the relative ease of working with conditional distributions. Given that the scale of the unobserved risk is arbitrary, I define the mean and variance of p_i to be zero and one. This is only done for notational simplicity and p_i can be redefined such that is has mean zero and variance one.

We are now ready to state the main result of this section.

Proposition 2. *When test data are selected based on the classifier that we want to evaluate, sensitivity is lower for all points on the interior of the ROC curve.*

The assumed bivariate distribution is a sufficient but not necessary condition for Proposition 2. The downward bias in the ROC curve is created by truncating the distribution of the classifier's output. Truncation causes an attenuation bias in perceived correlation between the classifier's output and the latent propensity to be a positive case. This attenuation bias causes the AUC to be smaller. Fig. 1 illustrates the effect of truncation on the probability density function (pdf) for the classifier's output.

There is an important feature of this setup to note before moving to the next section. Given this probabilistic formulation, both sensitivity and specificity are affected by p^* as long as a and p are not independent. This means that the skew of the test data will affect the ROC curve. (This type of effect has been discussed by Webb and Ting [23] and Fawcett and Flach [9].) Specifically, keeping ρ_{ap} constant, the AUC is increasing the absolute value of p^*. This is related to the reported tendency of ROC curves to be "overly optimistic" when the data are skewed ([5], p. 233 and [10], p. 79).[4]

[4] This is surprising because an advantage of ROC curves is that they can be invariant to class skew (see, for example, [20], p. 26 and Fawcett (2006), p. 864). Webb and Ting [23] explain that ROC curves are only invariant to class skew when we can think of features as coming from difference distributions for positive and negative cases.

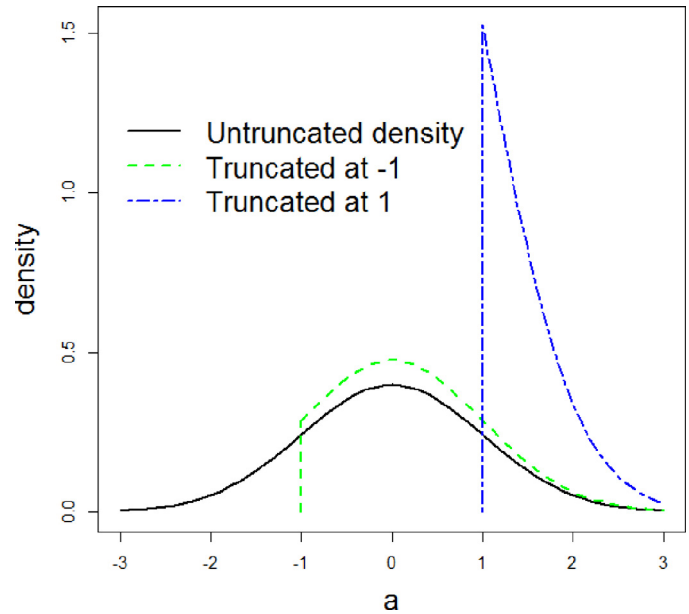

Fig. 1. The solid line is the pdf for the classifier's output. The dashed and dot-dash lines are the pdfs that result from truncating the classifier's output at −1 and 1, respectively.

3. ROC curves for nonrandom test data

This paper's procedure for creating ROC curves that are robust to sample selection is to infer the predictive power of the classifier (taking selection into consideration), then draw the ROC curve that is implied by our distributional assumptions. The proposed procedure has the following three steps.

Step 1. Subtract the mean and divide by the standard deviation to standardize the classifier's output. The mean and standard deviation should be based on all of the data, not only on the selected instances.

Step 2. Estimate p^* and the correlation between the classifier's output and the latent propensity to be a positive case, i.e. ρ_{ap}, by maximizing the following likelihood function with respect to the parameters p^*, ρ_{ap}, $\rho_{\varepsilon p}$, δ, γ:

$$\begin{aligned} L = \prod_i & \Phi_2(W, -Q\,;\rho_{\varepsilon p})^{\mathbb{1}(\text{outcome}_i=\text{positive})} \\ & \times \Phi_2(W, Q\,;-\rho_{\varepsilon p})^{\mathbb{1}(\text{outcome}_i=\text{negative})} \\ & \times \Phi(-W)^{\mathbb{1}(\text{outcome}_i=\text{NA})}, \end{aligned} \tag{11}$$

where

$$W = \delta X_i + \gamma\, a_i - s,$$

$$Q = (p^* - a_i\, \rho_{ap})/\sqrt{1-\rho_{ap}^2},$$

and $\mathbb{1}(\cdot)$ is the indicator function.

Step 3. Draw the ROC curve that is implied by our estimates in Step 2 and

$$\text{Sensitivity}(\text{ Specificity}) = \text{Prob}(a_i > c \mid p_i > p^*),$$

where c satisfies

$$\text{Specificity} = \text{Prob}(a_i < c \mid p_i < p^*).$$

To draw the ROC implied by these estimates (denoted here as $\widehat{p^*}$ and $\widehat{\rho_{ap}}$), begin with a set of cutoffs with sufficiently large range (e.g., −4–4). For each cutoff $c \in [-4, 4]$, we find the corresponding value of sensitivity as

$$\text{Prob}(a_i > c \mid p_i > p^*) = \left[1 - \Phi(\widehat{p^*})\right]^{-1} \int_c^{\infty} \phi(a) \left[1 - \Phi\left(\left[\widehat{p^*} - \widehat{\rho_{ap}}\, a\right] / \sqrt{1 - \widehat{\rho_{ap}}^2}\right)\right] da \tag{12}$$

and specificity as

$$\text{Prob}(a_i < c \mid p_i < p^*) = \Phi(\widehat{p^*})^{-1} \int_{-\infty}^{c} \phi(a)\, \Phi\left(\left[\widehat{p^*} - \widehat{\rho_{ap}}\, a\right] / \sqrt{1 - \widehat{\rho_{ap}}^2}\right) da. \tag{13}$$

The likelihood function in Step 2 is a reparameterization of the likelihood derived by Van de Ven and Van Pragg [21]. The ROC curve that we draw in Step 3 is a deterministic function of the maximum likelihood estimates from Step 2. By the functional invariance property of maximum likelihood estimates, we know that the ROC curve drawn in Step 3 is a consistent estimate of the expected ROC curve.

4. An example with wine-quality data

To provide a demonstration of this procedure with non-simulated data, I use data on white wine quality from Cortez et al. [3].[5] This dataset contains eleven attributes for 4898 white wines, including alcohol content, citric acid, and residual sugar. A detailed description of this data are provided by Cortez et al. For the measure of wine quality, each wine was evaluated by experts and given a score from zero to ten (with ten being the highest quality). Because we are interested in binary prediction, I define a wine with a score of six or higher as "good wine" and other wines as "not good wine."

All eleven attributes are used as predictors in a random forest classifier (based on Breiman [2] and implemented in R using Liaw and Wiener's [16] randomForest package) to predict (binary) wine quality. The random forest contains 1000 trees and tries three attributes at each split. I use the first 2/3 of the observations (3233 observations) as training data and the remaining 1/3 (1665 observations) as test data.

I first find the ROC curve for the random forest classifier using the full set of test data. The area under the ROC curve is .83. Next, let us suppose that the wine experts do not have enough time to score all of the wine in the test data. Preferring to taste wine that is more likely to be good wine, the experts taste the half the test data that the random forest classifier predicted was most likely to be good wine. With only half of the test data available, the area under the ROC curve falls to .60.

I now perform the procedure described in the previous section with the half of the test data predicted by be most likely to be good wine. I standardize the random forest scores and follow Step 2 to find the estimates $\widehat{\rho_{ap}} = .64$ and $\widehat{p^*} = -.55$.

Fig. 2 plots the ROC curves that are obtained with the full set of test data, the half of the test data that received a high score from the random forest, and the ROC curve based on our estimates of p^* and ρ_{ap}. The ROC curve based on our estimates of p^* and ρ_{ap} closely matches the ROC curve obtained with the full set of test data.

5. Simulation

This section reports the results of simulation exercises for the procedure presented in Section 3. The purpose of these simulations is to illustrate the performance of inferred ROC curves as well as the bias that arises in empirical ROC curves.[6]

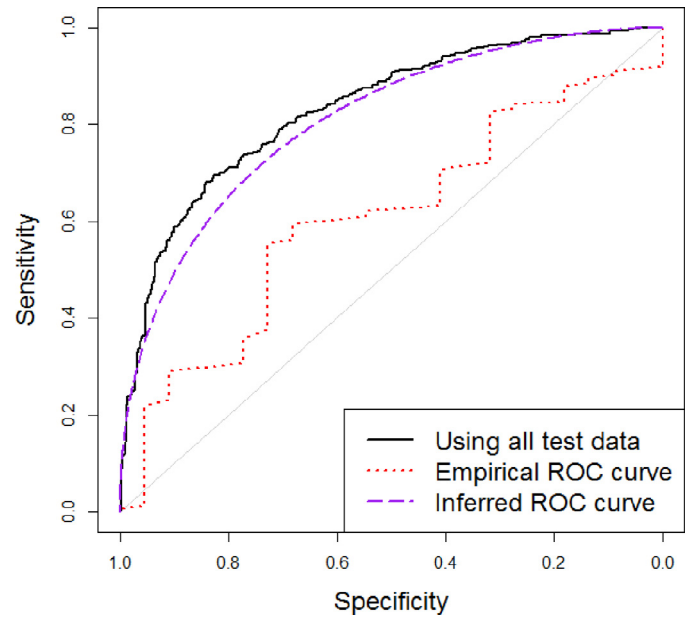

Fig. 2. ROC curves for wine quality prediction, as described in Section 5. The area under the ROC curve that uses all of the test data is .83. The areas under the empirical and inferred ROC curves, which only use half of the test data, are .60 and .81, respectively.

I first simulate the ROC curve that is obtained with random test data. For each Monte Carlo run, I draw 500 observations from the distribution

$$\begin{pmatrix} p_i \\ a_i \\ \varepsilon_i \end{pmatrix} \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} & \rho_{\varepsilon p} \\ \rho_{ap} & 1 & 0 \\ \rho_{\varepsilon p} & 0 & 1 \end{bmatrix}\right) \tag{14}$$

and define the outcome as in Eq. (3). I then calculate the AUC for $\mathcal{A}$. This serves as an estimate of the unbiased AUC.

Next, I simulate the ROC curve that results with selection and with the inferred ROC curve. I draw 1000 observations from the distribution in Eq. (14) with the outcome defined as before. I then sort the values by $(\gamma a_i + \varepsilon_i)$ and keep the 500 largest. (This is equivalent to setting $\delta = 0$ in Eq. (6).) This is done so that there is a fair degree of selectivity, but the number of observations used to generate the ROC curve is the same for both the random and nonrandom samples. I then use the 500 nonrandom observations to calculate the AUC for $\mathcal{A}$ using a empirical ROC curve and using the previous section's procedure.

Across simulations, I vary the correlation between the classifiers and with the latent propensity to be a positive case. The correlation between the classifiers is

$$\text{Cor}(a_i, b_i) = \frac{\gamma}{1 + \gamma^2}$$

and the correlation between the classifier that was used to select the test data and the latent propensity to be a positive case is

$$\text{Cor}(p_i, b_i) = \frac{\gamma\, \rho_{ap} + \rho_{\varepsilon p}}{1 + \gamma^2}.$$

Table 2 provides empirical and inferred AUCs for two classifiers: the first with $\rho_{ap} = .2$, which I call the weak classifier, and the second with $\rho_{ap} = .7$, which I call the strong classifier. The values in the first row of Table 2, .590 and .830, provide

[5] These data are available at the University of California at Irvine's Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html.

[6] By "empirical ROC curves," I mean the ROC curve constructed using the algorithm described in [8]. All R code that was used to produce these simulations (as well as the example in Section 4) are available upon request.

Table 2
Results based on 10,000 simulations. Each simulation is based on a sample of 500 draws. Mean values are presented with standard deviations in parenthesis. The parameter p^* is set to zero for all simulations.

Evaluating a classifier with test data selected by an unobserved classifier	$\rho_{ap} = .2$ (Weak)	$\rho_{ap} = .7$ (Strong)
AUC for ROC curves with a random sample	.590 (.026)	.830 (.018)
For no correlation between classifiers and data selected by a strong classifier, i.e. $\gamma = 0,\ \rho_{\varepsilon p} = .7$		
AUC for empirical ROC curves with data selected by a classifier	.619 (.028)	.936 (.011)
AUC for inferred ROC curves with data selected by a classifier	.591 (.031)	.854 (.042)
For .5 correlation between classifiers and data selected by a strong classifier, i.e. $\gamma = 1,\ \rho_{\varepsilon p} = .7$		
AUC for empirical ROC curves with data selected by a classifier	.578 (.028)	.683 (.059)
AUC for inferred ROC curves with data selected by a classifier	.579 (.084)	.779 (.095)
For .5 correlation between classifiers and data selected by a weak classifier, i.e. $\gamma = 1,\ \rho_{\varepsilon p} = 0$		
AUC for empirical ROC curves with data selected by a classifier	.575 (.026)	.796 (.021)
AUC for inferred ROC curves with data selected by a classifier	.591 (.093)	.832 (.057)

the (unbiased) AUCs that are obtained with a random sample for the weak and strong classifiers. When there is no correlation between the classifiers ($\gamma = 0$), there is an upward bias in the AUC for empirical ROC curves. As discussed at the end of Section 2, given our specification, AUC is increasing in the class skew. When $\rho_{ap} = \rho_{\varepsilon p} = .7$, the AUC for the empirical ROC curve with selection is .936. This is a 13% increase over the unbiased AUC of .830.

When there is a .5 correlation between the classifiers (i.e. $\gamma = 1$), the ROC curve is biased downward. For small positive values of γ (results not reported), there is an upward bias in ROC curves. As the correlation between the classifiers increases, the bias becomes more similar to the truncation bias described in Proposition 2. The simulations for which $\rho_{\varepsilon p} = 0$ illustrate a difference between the problem considered by this paper and the econometric literature on sample-selection bias. For a regression, when there is no correlation between the stochastic element in the selection equation and the stochastic element in the outcome equation, there is no bias. By contrast, Table 2 shows that there is a downward bias for ROC curves when $\rho_{\varepsilon p} = 0$. The AUCs for the weak and strong classifier, .575 and .796, are smaller than the unbiased values of .590 and .830, respectively.

For a .5 correlation between classifiers and data selected by a strong classifier ($\gamma = 1$ and $\rho_{\varepsilon p} = .7$), there is a noticeable difference between the areas under the inferred ROC curves and the unbiased AUCs. The AUCs for the weak and strong classifiers' inferred ROC curves are .578 and .683, respectively. In results not shown, the differences between areas under the inferred ROC curves and the unbiased AUCs are decreasing in the sample size. Appendix B shows that, when the output of classifier that selected the test data (b_i) is observed, the average AUCs from inferred ROC curves are much closer to the unbiased AUCs.

6. Discussion and conclusion

When test data are selected by a classifier, the bias in the ROC curve can be difficult to sign. A ROC curve that is constructed with data that was selected by the classifier will be biased downward. For a sample-selection procedure that is only weakly correlated with classifier being evaluated, the ROC curve may be biased upward. This paper presents a procedure for creating ROC curves that provide a consistent estimate of the ROC curve that would be obtained with random test data.

The procedure introduced here assumes that the classifier's output and latent propensity to be a positive case follow a bivariate normal distribution. The bivariate normal distribution has a relatively simple functional form. If non-Gaussian distributions are preferred for classifiers' output, the multivariate distributions used in this paper could be written in terms of copulas, as has been done for sample-selection bias in a regression setting (as in [15]). Copulas provide a way for handling bivariate distributions that result from arbitrary combinations of distributions. Copulas could be used to, for example, specify a Pareto distribution for the classifier's output while maintaining the normal distribution for the latent propensity to be a positive case. Alternatively, one could discretize the classifier's output and treat a as a latent variable (in the same manner as p).

An advantage of these distributional assumptions is that the estimation of the parameter p^* leads to an estimate of the percent of positive cases in the population. This parameter may be of interest to an organization like the IRS that could use $\widehat{p^*}$ to estimate the percent of tax returns that contain errors. Given $\widehat{p^*}$, a consistent estimate of the portion of tax returns that contain errors is provided by $\Phi(\widehat{p^*})$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF).

This paper's results have implications for the practice of sample enrichment, which typically involves removing cases from the data [14]. In addition to possible effects from changes to the distribution of positive and negative cases (as discussed at the end of Section 2), if the way in which cases are removed is correlated with the cases' propensity to be positive or with the classifier's output, the empirical ROC curves constructed with that sample will be biased. A simple test for this bias is to perform the procedure Section 3 and test whether $\rho_{\varepsilon p}$ and γ both equal zero.

Acknowledgments

I thank the editor and reviewers for helpful comments. I am also grateful for comments from Mike Gurbutt, Richard Hahn, Chris Hansen, Daniel Jessie, Patricia Ledesma, Joon-Suk Lee, Christian Leuz, Phillip Li, Saad Siddiqui, and Steve Yang.

Appendix A. Proofs

Proof of Lemma 1. For (i):

We want to show that

$$\text{Prob}(a_i > c \mid p_i > p^*) < \text{Prob}(a_i > c \mid p_i > p^*, a_i > s).$$

We assume that $\text{Prob}(p_i > p^*)$ and $\text{Prob}(a_i > s)$ are both nonzero. If our selection rule s were negative enough, selection would have no impact on sensitivity:

$$\lim_{s \to -\infty} \text{Prob}(a_i > c \mid p_i > p^*, a_i > s) = \text{Prob}(a_i > c \mid p_i > p^*).$$

We will now show that sensitivity is monotonically increasing in the selection rule s. We first rewrite specificity in terms of the pdf of a_i conditional on $(p_i > p^*)$ as

$$\text{Prob}(a_i > c \mid p_i > p^*, a_i > s) = \frac{\text{Prob}(a_i > c \mid p_i > p^*)}{\text{Prob}(a_i > s \mid p_i > p^*)} = \frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\int_s^\infty f_{a|p>p^*}(a_i)da_i},$$

where $f_{a|p>p^*}$ is pdf of a_i conditional on $(p_i > p^*)$. We take the derivative of specificity conditional on selection with respect to s through a straight-forward application of the Leibniz rule:

$$\frac{d}{ds}\left(\frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\int_s^\infty f_{a|p>p^*}(a_i)da_i}\right) = f_{a|p>p^*}(s)\frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2} > 0.$$

For (ii):

We want to show that

$$\text{Prob}(a_i < c \mid p_i < p^*) > \text{Prob}(a_i < c \mid p_i < p^*, a_i > s).$$

We assume that $\text{Prob}(p_i < p^*)$ and $\text{Prob}(a_i > s)$ are both nonzero. As in part (i), we begin by noting that if our selection rule s were negative enough, selection would have no impact on specificity:

$$\lim_{s\to-\infty} \text{Prob}(a_i < c \mid p_i < p^*, a_i > s) = \text{Prob}(a_i < c \mid p_i < p^*).$$

We will now show that specificity is monotonically decreasing in the selection rule s. Again, we first rewrite specificity in terms of the pdf of a_i conditional on $(p_i < p^*)$ as

$$\begin{aligned}\text{Prob}(a_i < c \mid p_i < p^*, a_i > s) &= \frac{\text{Prob}(s < a_i < c \mid p_i < p^*)}{\text{Prob}(a_i > s \mid p_i < p^*)} \\ &= \frac{\int_s^c f_{a|p<p^*}(a_i)da_i}{\int_s^\infty f_{a|p<p^*}(a_i)da_i},\end{aligned}$$

where $f_{a|p<p^*}$ is pdf of a_i conditional on $(p_i < p^*)$. We take the derivative of specificity conditional on selection with respect to s by applying the Leibniz rule:

$$\frac{d}{ds}\left(\frac{\int_s^c f_{a|p<p^*}(a_i)da_i}{\int_s^\infty f_{a|p<p^*}(a_i)da_i}\right) = -\frac{f_{a|p<p^*}(s)\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i - \int_s^c f_{a|p<p^*}(a_i)da_i\right]}{\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i\right]^2} < 0.$$

□

Proof of Proposition 2. Here, I show that sensitivity for a given level of specificity is a decreasing function of s. Since this term approaches a point on the ROC curve as s approaches negative infinity, a monotonic decrease in s implies that any point on the ROC curve will be lower.

I define sensitivity for a given level of specificity and selection rule s as

$$\text{Sensitivity}(\text{ Specificity}, s\ ) = \text{Prob}(a_i > c \mid p_i > p^*, a_i > s),$$

where c satisfies

$$\text{Specificity} = \text{Prob}(a_i < c \mid p_i < p^*, a_i > s),$$

assuming that $\text{Prob}(p_i < p^*)$, $\text{Prob}(p_i > p^*)$, and $\text{Prob}(a_i > s)$ are all nonzero. For a fixed level of specificity, the effect of an increase in s on sensitivity is

$$\frac{d\,\text{Sensitivity}}{d\,s} = \underbrace{\frac{\partial\,\text{Sensitivity}}{\partial\,s}}_{\text{Direct effect of } s \text{ on sensitivity}} + \underbrace{\frac{\partial\,\text{Sensitivity}}{\partial\,c}\frac{d\,c}{d\,s}}_{\text{Indirect effect of changing } c}.$$

These terms are

$$\frac{\partial\,\text{Sensitivity}}{\partial\,s} = f_{a|p>p^*}(s)\frac{\int_c^\infty f_{a|p>p^*}(a_i)da_i}{\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2} > 0,$$

$$\frac{\partial\,\text{Sensitivity}}{\partial\,c} = -\frac{f_{a|p>p^*}(c)}{\int_s^\infty f_{a|p>p^*}(a_i)da_i} < 0, \quad \text{and}$$

$$\begin{aligned}\frac{d\,c}{d\,s} &= -\frac{\partial\,\text{Prob}(a_i < c \mid p_i < p^*, a_i > s)/\partial\,c}{\partial\,\text{Prob}(a_i < c \mid p_i < p^*, a_i > s)/\partial\,s} \\ &= \frac{f_{a|p<p^*}(c)\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i\right]}{f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]} > 0,\end{aligned}$$

where the last term follows from the use of the implicit function theorem. After applying some high-school algebra, $d\,\text{Sensitivity}/d\,s$ can be written as

$$\begin{aligned}\frac{d\,\text{Sensitivity}}{d\,s} &= \frac{f_{a|p>p^*}(s)f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p>p^*}(a_i)da_i\right]\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]}{f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2} \\ &\quad - \frac{f_{a|p>p^*}(c)f_{a|p<p^*}(c)\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]\left[\int_s^\infty f_{a|p<p^*}(a_i)da_i\right]}{f_{a|p<p^*}(s)\left[\int_c^\infty f_{a|p<p^*}(a_i)da_i\right]\left[\int_s^\infty f_{a|p>p^*}(a_i)da_i\right]^2}.\end{aligned}$$

The denominator is clearly positive so we focus on the numerator. Given the bivariate distribution that we assumed, the condition for the numerator to be negative is

$$\begin{aligned}&[\phi(s)]^2\Phi(-\rho_{ap}s/(1-\rho_{ap})^2)\Phi(\rho_{ap}s/(1-\rho_{ap})^2) \\ &\times\int_c^\infty \phi(a_i)\Phi(-a_i\rho_{ap}/(1-\rho_{ap}^2))da_i \\ &\times\int_c^\infty \phi(a_i)\Phi(a_i\rho_{ap}/(1-\rho_{ap}^2))da_i \\ &< [\phi(c)]^2\Phi(-\rho_{ap}c/(1-\rho_{ap})^2)\Phi(\rho_{ap}c/(1-\rho_{ap})^2) \\ &\times\int_s^\infty \phi(a_i)\Phi(-a_i\rho_{ap}/(1-\rho_{ap}^2))da_i \\ &\times\int_s^\infty \phi(a_i)\Phi(a_i\rho_{ap}/(1-\rho_{ap}^2))da_i.\end{aligned}$$

This condition holds for $c > s$. It follows that

$$\frac{d\,\text{Sensitivity}}{d\,s} < 0,$$

which implies that, for a fixed level of specificity, sensitivity is monotonically decreasing in selectivity s. □

Appendix B. Evaluating a classifier with test data selected by an observed classifier

As in the main text, I consider the case of selection of test data based another classifier, b. Instance i is selected if $b_i > s$ and not selected otherwise. Unlike the main text, this section explores situations in which b is observed. As before, I allow for correlation between the output of classifiers $\mathcal{A}$ and $\mathcal{B}$, which could arise from using similar attributes to make predictions:

$$\begin{pmatrix} p_i \\ a_i \\ b_i \end{pmatrix} \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ap} & \rho_{bp} \\ \rho_{ap} & 1 & \rho_{ab} \\ \rho_{bp} & \rho_{ab} & 1 \end{bmatrix}\right).$$

The likelihood function for the data can be expressed as

$$\begin{aligned}L = \prod_i &\Phi(-[p^* - E(p_i|a_i, b_i)]/\sigma_{p|ab})^{\mathbb{1}(\text{outcome}_i=\text{positive})} \\ &\times\Phi([p^* - E(p_i|a_i, b_i)]/\sigma_{p|ab})^{\mathbb{1}(\text{outcome}_i=\text{negative})} \times \phi(a_i, b_i),\end{aligned} \tag{B.1}$$

where $\sigma_{p|ab}$ is the standard deviation of p conditional on a and b,

$$\sigma_{p|ab} \equiv \sqrt{1 - \frac{1}{1-\rho_{ab}^2}\left[(\rho_{ap} - \rho_{bp}\rho_{ab})\rho_{ap} + (\rho_{bp} - \rho_{ap}\rho_{ab})\rho_{bp}\right]},$$

and the expectation of p_i conditional on a_i and b_i is

$$E(p_i|a_i, b_i) = \frac{1}{1-\rho_{ab}^2}\left[(\rho_{ap} - \rho_{bp}\rho_{ab})a_i + (\rho_{bp} - \rho_{ap}\rho_{ab})b_i\right].$$

We can estimate the parameters ρ_{ap}, ρ_{ab}, ρ_{bp}, and p^* by maximizing the likelihood function in Eq. (B.1).

Table A3
Results based on 10,000 simulations. Each simulation is based on a sample of 500 draws. Mean values are presented with standard deviations in parenthesis. The parameter p^* is set to zero for all simulations.

Evaluating a classifier with test data selected by an observed classifier	$\rho_{ap} = .2$ (Weak)	$\rho_{ap} = .7$ (Strong)
AUC for ROC curves with a random sample	.590 (.026)	.830 (.018)
For no correlation between classifiers and data selected by a strong classifier, i.e. $\rho_{ab} = 0,\ \rho_{bp} = .7$		
AUC for empirical ROC curves with data selected by a classifier	.619 (.029)	.936 (.011)
AUC for inferred ROC curves with data selected by a classifier	.590 (.023)	.829 (.015)
For .5 correlation between classifiers and data selected by a strong classifier, i.e. $\rho_{ab} = .5,\ \rho_{bp} = .7$		
AUC for empirical ROC curves with data selected by a classifier	.533 (.027)	.800 (.021)
AUC for inferred ROC curves with data selected by a classifier	.590 (.025)	.829 (.015)
For .5 correlation between classifiers and data selected by a weak classifier, i.e. $\rho_{ab} = .5,\ \rho_{bp} = .2$		
AUC for empirical ROC curves with data selected by a classifier	.568 (.026)	.832 (.018)
AUC for inferred ROC curves with data selected by a classifier	.591 (.028)	.830 (.020)
For .5 correlation between classifiers and data selected by a nonpredictive classifier, i.e. $\rho_{ab} = .5,\ \rho_{bp} = 0$		
AUC for empirical ROC curves with data selected by a classifier	.599 (.025)	.863 (.016)
AUC for inferred ROC curves with data selected by a classifier	.591 (.029)	.830 (.021)

As in the main text, I use a simulation study to examine the performance of the procedure. Table A.3 presents these results. Not surprisingly, when the classifier that selected the test data is observed, the average areas under the inferred ROC curves are much closer to the average areas under the ROC curves that are based on random samples. We also see that the standard deviations of the areas under the inferred ROC curves are closer to the standard deviations of the areas under the ROC curves based on random samples. The average AUCs for the inferred ROC curves are nearly indistinguishable from their unbiased values of .590 and .830.

References

[1] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159.
[2] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decis. Support Syst. 47 (4) (2009) 547–553.
[4] J. Crook, J. Banasik, Does reject inference really improve the performance of application scoring models? J. Bank Financ. 28 (4) (2004) 857–874.
[5] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 233–240.
[6] D.D. Dorfman, E. Alf, Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals: rating-method data, J. Math. Psychol. 6 (3) (1969) 487–496.
[7] A. Erkanli, M. Sung, E. Jane Costello, A. Angold, Bayesian semi-parametric ROC analysis, Stat. Med. 25 (22) (2006) 3905–3928.
[8] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[9] T. Fawcett, P.A. Flach, A response to webb and tings on the application of roc analysis to predict classification performance under varying class distributions, Mach. Learn. 58 (1) (2005) 33–38.
[10] H. He, Self-adaptive Systems for Machine Intelligence, John Wiley & Sons, 2011.
[11] J.J. Heckman, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, Ann. Econ. Social Meas. 5 (4) (1976) 475–492.
[12] J.J. Heckman, Sample selection bias as a specification error, Econometrica 47 (1) (1979) 153–161.
[13] J. Hernández-Orallo, P. Flach, C. Ferri, ROC curves in cost space, Mach. Learn. 93 (1) (2013) 71–91.
[14] P.D. Leber, C.S. Davis, Threats to the validity of clinical trials employing enrichment strategies for sample selection, Control. Clin. Trials 19 (2) (1998) 178–187.
[15] P. Li, M.A. Rahman, Bayesian analysis of multivariate sample selection models using Gaussian copulas, Adv. Econom. 27 (2011) 269.
[16] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22.
[17] C.E. Metz, X. Pan, “Proper” binormal ROC curves: theory and maximum-likelihood estimation, J. Math. Psychol. 43 (1) (1999) 1–33.
[18] A. Moffat, Seven numeric properties of effectiveness metrics, in: Information Retrieval Technology, Springer, 2013, pp. 1–12.
[19] D.J. Poirier, Partial observability in bivariate probit models, J. Econom. 12 (2) (1980) 209–217.
[20] J.A. Swets, R.M. Dawes, J. Monahan, Psychological science can improve diagnostic decisions, Psychol. Sci. Public Interest 1 (1) (2000) 1–26.
[21] W.P.M.M. Van de Ven, B.M.S. Van Praag, The demand for deductibles in private health insurance: a probit model with sample selection, J. Econom. 17 (2) (1981) 229–252.
[22] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, A theoretical analysis of NDCG type ranking measures, in: Proceedings of the 26th Annual Conference on Learning Theory, 2013.
[23] G.I. Webb, K.M. Ting, On the application of ROC analysis to predict classification performance under varying class distributions, Mach. Learn. 58 (1) (2005) 25–32.
[24] K.H. Zou, W. Hall, Two transformation models for estimating an ROC curve derived from continuous data, J. Appl. Stat. 27 (5) (2000) 621–631.