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.

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/

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*