Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Review Article
Volume 8 Issue 1 - August 2018
DOI: 10.19080/BBOAJ.2018.07.555728

Biostat Biometrics Open Acc J

Metrics Used When Evaluating the Performance of Statistical Classifiers

Daniel R Jeske*

Department of Statistics, University of California, USA

Submission: June 05, 2018; **Published:** August 01, 2018

***Corresponding author:** Daniel R Jeske, Department of Statistics, University of California, Riverside, CA, USA, Tel: 951-827-3014; Email: daniel.jeske@ucr.edu

Abstract

This article reviews important performance metrics that are used to evaluate the accuracy of statistical classifiers. How the metrics are used to construct Receiver Operator Characteristic (ROC) curves, Predictive ROC (PROC) curves, and Precision-Recall (PR) curves is also discussed. Relationships between the metrics are revealed.

Keywords: False positive rate; False negative rate; Specificity; Sensitivity; Positive predictive value; Negative predictive value; Precision; Recall; Youden threshold

Abbreviations: ROC: Receiver Operator Characteristic Curves; PROC: Predictive ROC Curves; PR: Precision-Recall; AUC: Area Under the Curve; NPV: Negative Predictive Value; PPV: Positive Predictive Value; FPR: False Positive Rate; FNR: False Negative Rate

Introduction

A statistical classifier maps a set of features X, to a class variable C. The features X can be a mix of categorical and interval variables and the class C is one of a finite number of possible classes. Applications frequently are concerned with two classes, and in this context the classifier is referred to as a binary classifier. In medical diagnostic applications, X could represent patient characteristics and $C=0$ $(C=1)$ might correspond to healthy (diseased) patient status.

There are a number of methods available for developing a statistical classifier, including Bayes classifiers, tree classifiers, support vector classifiers, neural network classifiers, logistic regression classifiers, and ensemble classification methods. See, for example, reference [1], for details on these methods. Using training data that has both features and the class label for a sample of subjects, the classification methods construct a predictive function $T(\cdot)$ that maps X to a predicted class label, $\hat{C}$. For binary classifiers,

$$\hat{C}=\begin{cases}1 & \text{, if } T(X)\geq u\\ 0 & \text{, if } T(X)<u\end{cases} \qquad (1)$$

where u is a threshold that is determined to trade-off performance objectives for the classifier. Equation (1) assumes, without loss of generality, that large values of $T(X)$ correlate to class $C=1$. It is understood in practice that no single classification method works uniformly the best, and typically investigators will experiment with a variety of options and choose the one the works best for their application.

When choosing the threshold u, there are a four important performance metrics that should be examined. The key to understanding these metrics is the notion of class-conditional distributions of $T(X)$. Let $F_0(F_1)$ denote the conditional cumulative distribution functions of $T(X)$ given $C=0$ $(C=1)$.

The ROC curves

The first two performance metrics of importance are the false positive rate (FPR) and false negative rate (FNR), defined as

$$\begin{aligned}FPR(u)&=P(\hat{C}=1\,|\,C=0)=1-F_0(u)\\ FNR(u)&=P(\hat{C}=0\,|\,C=1)=F_1(u).\end{aligned} \qquad (2)$$

Alternative terminology used with the ROC curve is *sensitivity* and *specificity*, which are defined as

$$\begin{aligned}specificity\,(u)&=1-FPR\,(u)\\ sensitivity\,(u)&=1-FNR\,(u).\end{aligned} \qquad (3)$$

The Receiver Operating Characteristic (ROC) curve is a plot of the locus of points defined by $(1-specificity\,(u),\ sensitivity\,(u))=(1-F_0(u),1-F_1(u))$, obtained by varying u. Figure 1 shows a schematic picture of an ROC curve, and it can be seen how it facilitates choosing a threshold u that strikes a balance between the conflicting objectives of simultaneously achieving high sensitivity and high specificity [2-4]. A commonly used threshold is $u^*=\underset{u}{\operatorname{argmin}}(FPR(u)+FNR(u))$ which is known as the Youden threshold [5]. An alternative threshold is the point on the ROC curve that is closest to the optimal point $(0,1)$.

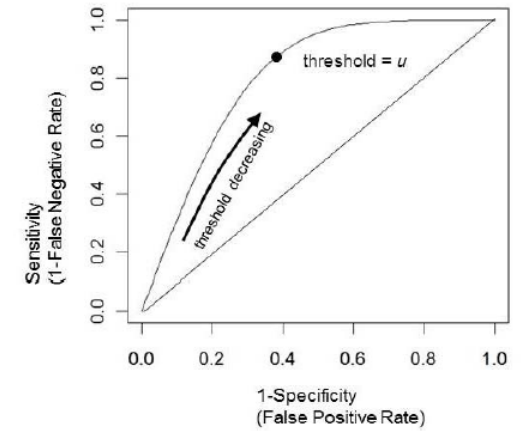

Figure 1: Schematic Diagram of an ROC Curve.

The area under the curve (AUC) is often reported as a measure of merit for the particular methodology used to develop the classifier [6]. AUC is a global measure that is not particular to a single threshold, and as such it loses its relevance with a specific implementation of the classifier that requires choosing one threshold.

The PROC Curve

A second pair of important performance metrics for a classifier are negative predictive value (NPV) and positive predictive value (PPV), defined as

$$NPV(u) = P(C = 0 \mid \hat{C} = 0)$$
$$PPV(u) = P(C = 1 \mid \hat{C} = 1). \quad (4)$$

NPV and PPV have the interpretation of the fraction of class 0 predictions that are correct and the fraction of class 1 predictions that are correct, respectively. Whereas FPR and FNR measure error rates of the classifier before the prediction is made (a-priori), NPV and PPV measure the accuracy of the classifier after the prediction is made (a-posteriori). In medical diagnostic applications, FPR and FNR aid in determining whether or not it is useful to perform the diagnostic procedure and NPV and PPV aid in interpreting the results if it is performed. Each of the metrics plays a role in providing a comprehensive assessment of the performance capability of the classifier.

In order to calculate NPV and PPV, it is necessary to know the prevalence of class $C = 1$, denoted by π_1 This necessity is revealed in the following formulas for NPV and PPV which follow from use of Bayes' rule,

$$NPV(u) = \frac{(1 - FPR(u))(1 - \pi_1)}{(1 - FPR(u))(1 - \pi_1) + FNR(u)\pi_1} = \frac{F_0(u)(1 - \pi_1)}{F_0(u)(1 - \pi_1) + F_1(u)\pi_1} \quad (5)$$

$$PPV(u) = \frac{(1 - FNR(u))\pi_1}{(1 - FNR(u))\pi_1 + FPR(u)(1 - \pi_1)} = \frac{(1 - F_1(u))\pi_1}{(1 - F_1(u))\pi_1 + (1 - F_0(u))(1 - \pi_1)}.$$

The predictive ROC (PROC) curve is a plot of the locus of points $(1 - NPV(u), PPV(u))$, obtained by varying u. Unlike the ROC curve, which is always monotone increasing, the PROC curve need not be monotone increasing. Monotonicity of the PROC curve requires the hazard and reversed hazard functions of F_0 and F_1 be ordered [7]. Figure 2 illustrates the general result that when F_0 and F_1 are homogenous normal distributions, the PROC curve is monotone, but it is not monotone for the heterogeneous normal case.

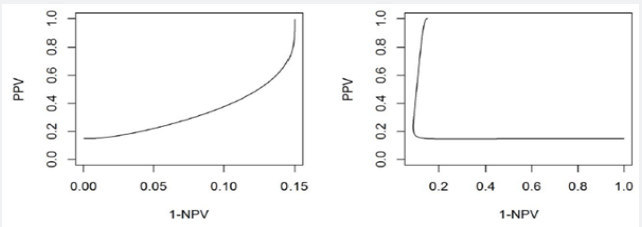

Figure 2: PROC curves for two cases where distribution of $T(X)$ is a normal distribution. Left panel has $F_0 \sim N(0,1)$ and $F_1 \sim N(1,1)$. . Right panel has $F_0 \sim N(0,1)$ and $F_1 \sim N(1,\sqrt{3})$.

Discussion

The literature on classifier performance metrics also includes discussion of the precision-recall (PR) curve [8-9]. Precision is an alternative term for PPV and recall is an alternative term for sensitivity. The PR curve is therefore an alternative plot for showing two of the four important performance metrics that have been discussed. The diversity in the references included in this review reflect the fact that research pertaining to the use and evaluation of statistical classifiers span a variety of different disciplines.

References

1. James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning, Springer, New York.
2. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39(4): 561-577.
3. Fawcett T (2006) An Introduction to ROC Analysis. Pattern Recognition Letters 27(8): 861-874.
4. Baker S (2003) The Central Role of Receiver Operating Characteristic (ROC) Curves in Evaluating Tests for the Early Detection of Cancer. Journal of the National Cancer Institute 95: 511-515.
5. Youden WJ (1950) Index for Rating Diagnostic Tests. Cancer 3(1): 32-35.
6. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1): 29-36.
7. Shiu SY, Gatsonis C (2008) The Predictive Receiver Operating Characteristic Curve for the Joint Assessment of the Positive and Negative Predictive Values. Philos Trans A Math Phys Eng Sci 366(1874): 2313-2333.
8. Saito T, Rehmsmeier M (2015) The Precision-Recall Plot is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One 10(3): e0118432.
9. Davis J, Goadrich M (2006) The Relationship between Precision-Recall and ROC Curves, Proceedings of the 23rd International Conference on Machine Learning, Pp. 232-240.

How to cite this article: Daniel R Jeske. Metrics Used When Evaluating the Performance of Statistical Classifiers. Biostat Biometrics Open Acc J. 2018; 8(1): 555728. DOI: 10.19080/BBOAJ.2018.08.555728.

DOI: 10.19080/BBOAJ.2018.08.555728

How to cite this article: Daniel R Jeske. Metrics Used When Evaluating the Performance of Statistical Classifiers. Biostat Biometrics Open Acc J. 2018; 8(1): 555728. DOI: 10.19080/BBOAJ.2018.08.555728.