Computational Statistics and Data Analysis 93 (2016) 192–208

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Modelling receiver operating characteristic curves using Gaussian mixtures

Amay S.M. Cheam *, Paul D. McNicholas

Department of Mathematics and Statistics, McMaster University, ON, Canada

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form 20 April 2015
Accepted 22 April 2015
Available online 7 May 2015

Keywords:
Binormal curve
EM algorithm
Gaussian mixture distributions
LABROC
Mixture models
Monte Carlo method
ROC curve

ABSTRACT

The receiver operating characteristic (ROC) curve is widely applied in measuring the performance of diagnostic tests. Many direct and indirect approaches have been proposed for modelling the ROC curve and, because of its tractability, the Gaussian distribution has typically been used to model both diseased and non-diseased populations. Using a Gaussian mixture model leads to a more flexible approach that better accounts for atypical data. The Monte Carlo method can be used to circumvent the absence of a closed-form for a functional form of the ROC curve. The proposed method, in which a Gaussian mixture is used in conjunction with the Monte Carlo method, performs favourably when compared to the crude binormal curve and the semi-parametric frequentist binormal ROC using the well-known LABROC procedure.

1. Introduction

The receiver operating characteristic (ROC) curve has gained tremendous popularity since its use in signal detection theory during World War II. The necessity to evaluate performance of a diagnostic test, as noted by Lusted (1971), has resulted in the increased attention received by the ROC curve. In addition to being a useful tool for evaluating the efficiency of a diagnostic test, the ROC curve also presents a practical way to select an optimal threshold and to compare different tests. However, the empirical ROC curve is not desirable for the simple reason that it violates certain theoretical properties. Many authors have proposed different ways to model the ROC curve to circumvent this issue. Approaches to modelling the ROC curve within the literature can be divided into two categories: direct and indirect.

The direct approach, which is less appealing, does not depend on any distributional hypotheses. The idea is to construct the ROC curve directly from the population scores; in medical settings, these are often divided into two groups, diseased and non-diseased, without any assumptions (Lloyd, 1998; Zhou and Harezlak, 2002). As mentioned previously, the empirical ROC curve violates certain theoretical properties, e.g., it is not necessarily monotonically increasing. To overcome this obstacle, some authors have proposed non-parametric estimation of the density function of each population using kernel smoothing methods (Hall and Hyndman, 2003; Lloyd, 1998; López-de Ullibarri et al., 2008; Qiu and Le, 2001; Zou et al., 1997). Hence, the problem is reduced to selection of an optimal bandwidth (Lloyd, 1998; Peng and Zhou, 2004; Zhou and Harezlak, 2002). Lloyd (1998) suggests using the bootstrap to minimize any distortion when smoothing the ROC curve.

The indirect approach assumes that each population follows a certain distribution and implicitly derives a functional form for the ROC curve. Both parametric and semi-parametric methods have been proposed to construct the curve. One of the parametric methods assumes that diseased and non-diseased populations follow a family of distributions, such as Gaussian,

* Correspondence to: Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada, L8S 4L8. Tel.: +1 905 525 9140x23414.
E-mail addresses: cheamas@math.mcmaster.ca (A.S.M. Cheam), mcnicholas@math.mcmaster.ca (P.D. McNicholas).

http://dx.doi.org/10.1016/j.csda.2015.04.010
0167-9473/

which is the obvious and simple choice, the gamma (Dorfman et al., 1997), and others (Zweig and Campbell, 1993). Goddard and Hinberg (1990) point out that the Gaussian assumption is not always adequate in some scenarios, such as prostate cancer. The authors emphasize that an inconsiderate and careless application of the method is not recommended, because it depends strongly on distributional assumptions. Furthermore, Zhou et al. (2002) stress the need to carefully verify the consistency of data with the assumptions. An alternative is to specify a functional form of the ROC curve instead of assuming a distribution. For instance, both populations can be assumed to follow a logistic distribution with the same variance (Swets, 1986). England (1988) suggests an exponential model with two parameters. Both parametric methods are very similar because the distribution of the test scores entirely determines the shape of the ROC curve. The main advantages of a parametric method are simplicity, the smoothness of the curve, and an ability to work with a small number of parameters.

The semi-parametric method is more attractive in terms of flexibility due to the presence of non-parametric and parametric components. The binormal model (Green and Swets, 1966) is a good example; it assumes that both populations follow a Gaussian distribution after some monotonically increasing transformation (Hanley, 1996). Hence, the problem is reduced to estimating the parameters, i.e., the slope and intercept. A range of solutions has been proposed using different techniques, such as generalized least squares (Hsieh and Turnbull, 1996), maximum likelihood, pseudo-likelihood (Cai and Moskowitz, 2004; Zhou and Lin, 2008; Zou and Hall, 2000), and other methods. For example, to obtain a smooth binormal ROC curve, Metz et al. (1998) develop an algorithm called LABROC, which groups continuous data into a finite number of ordered categories and then uses the maximum likelihood algorithm from Dorfman and Alf (1968) for ordinal data. Li et al. (1999) suggest a variation of this method, where they model the scores of a diagnostic test for non-diseased and diseased patients non-parametrically and parametrically, respectively, with no functional relationship assumed between these two distributions. Instead of directly modelling the distributions of the diagnostic scores of the two populations when the true status of the disease is known, another approach is to model the probability of knowing the disease status of the diagnostic scores using logistic regression (Qin and Zhang, 2003). Like any estimation problem, lack-of-fit can be an issue for the semi-parametric method. In addition to this estimation problem, the construction of confidence bands, for a given choice of both population distributions, is complicated.

Our motivation is to develop a method that can give an estimate of the ROC curve with more flexibility and smoothness, produce reliable confidence bands, and ensure the natural monotonicity property of the ROC curve. We propose a Gaussian mixture (GM) distribution to model both non-diseased and diseased populations. This will enable us to capture more complex behaviour and distribution shapes than the traditional normality assumption. By combining the Monte Carlo method and the GM distribution, our method generates an ensemble of replica ROC curves and computes summary measures, such as the area under the curve (AUC), based on the ensemble.

The remainder of the paper is organized as follows. In Section 2, we provide some background on ROC curves, followed by details of our proposed approach (Section 3). Results from simulation studies are provided in Section 4 and real data analyses are discussed in Section 5. In Section 6, some concluding remarks are given and possible extensions are discussed.

2. Background

The ROC curve is defined as a plot of the true positive rate (TPR) against the false positive rate (FPR), or sensitivity versus 1−specificity, for various threshold values. This is generally a curve in the unit square anchored at (0, 0) and (1, 1), and above the line joining those points. Let $X \sim F$ and $Y \sim G$ be two independent continuous variables denoting the diagnostic test measure for non-diseased and diseased populations, respectively. By convention, a patient is considered diseased if the value of the score is greater than a specified threshold. Note that we borrow the notation of Gu et al. (2008) in some of what follows. For a given threshold value $c_t \in \mathbb{R}$,

$$\mathrm{FP}(c_t) = \int_{-\infty}^{+\infty} f_X(x) I\,(x - c_t)\,dx = P(X > c_t), \tag{1}$$

$$\mathrm{TP}(c_t) = \int_{-\infty}^{+\infty} g_Y(y) I\,(y - c_t)\,dy = P(Y > c_t), \tag{2}$$

where

$$I(u) = \begin{cases} 1, & \text{if } u > 0, \\ 0, & \text{if } u \leq 0. \end{cases}$$

Therefore, the ROC curve is obtained by

$$\{(t, R(t))\} = \{(\mathrm{FP}(c_t), \mathrm{TP}(c_t))\}, \tag{3}$$

where $t \in D \subset [0, 1]$.

When t is given, $c_t = \bar{F}^{-1}(t) = F^{-1}(1 - t)$, where $F^{-1}(\zeta) = \inf\{x : F(x) \geq \zeta\}$. If $\bar{F}^{-1}(t)$ exists, then the functional form of the ROC curve is given by

$$R(t) = TP(c_t) = \bar{G}(\bar{F}^{-1}(t)) = \bar{G}(c_t) = P(Y > c_t) = P(Y > \bar{F}^{-1}(t)), \tag{4}$$

where $\bar{F}(u) = P(X > u)$ and $\bar{G}(u) = P(Y > u)$ are known as survival functions of X and Y, respectively.