Tatra Mt. Math. Publ. **39** (2008), 215–223

TATRA MOUNTAINS
Mathematical Publications

TESTING THE DIFFERENCE OF THE ROC CURVES IN BIEXPONENTIAL MODEL

Martin Betinec

ABSTRACT. Receiver Operating Characteristics (ROC) curves are a useful instrument for evaluation of supervised classifiers performance. The classic form of ROC curve is applicable only to the two-state classification. Nevertheless, it covers a broad spectrum of applications (clinical diagnostics in medicine, computational linguistics, machine learning, data mining, etc.)

In this contribution we will derive a test of equivalence of two ROC curves in biexponential ROC model for unpaired design of data. The result will be illustrated on real data coming from a sociological research.

1. Introduction

The *Reciever Operating Characteristic* (ROC) curve is mainly used to assess the quality of a classifier in the following classification situation. Objects of interest are classified by the classifier γ either to the group $\mathcal{G}_1$ or to the group $\mathcal{G}_0$, where $\mathcal{G}_1 \cap \mathcal{G}_0 = \emptyset$. For simplicity of notation we identify each object with a vector of its features $\boldsymbol{X}$. The classifier γ makes *predictions* $\widehat{G}$ using the *score function* (sometimes called also *marker*, *diagnostic variable*, etc.) $Y = \gamma(\boldsymbol{X})$ and an arbitrary cut-point $\theta \in \mathbb{R}$, such as

$$\widehat{G} = \begin{cases} 1 & \text{if} \quad Y > \theta\,, \\ 0 & \text{if} \quad Y \le \theta\,. \end{cases} \tag{1}$$

Denote the distribution functions of Y conditional by the membership of objects as follows

$$F_0(y) = \mathsf{P}(Y \le \theta |\, \boldsymbol{X} \in \mathcal{G}_0) \quad \text{and} \quad F_1(y) = \mathsf{P}(Y \le \theta |\, \boldsymbol{X} \in \mathcal{G}_1)\,. \tag{2}$$

2000 Mathematics Subject Classification: Primary 62H30; Secondary 62F03, 91D99.
Keywords: ROC curves, classification, machine learning, hypotheses testing, parametrical models.
This research was supported by the GAČR grant No. 201/06/0186 of Czech Grant Agency and by research project MSM 0021620841.

Traditional measures of classifiers quality, such as the *misclassification error*, *accuracy* or the *risk,* fail when the prevalence of the $\mathcal{G}_1$-objects in population is small. In these situations it is necessary to evaluate the performance of the classifier separately on $\mathcal{G}_0$ and on $\mathcal{G}_1$. For $\theta \in \mathbb{R}$ define the *True Positive Rate* (TPR)—also called *Hit*, *Recall* or *Sensitivity*—and the *False Positive Rate* (FPR)—also called *Fallout, Alarm rate* or *Non-specificity*—as

$$\mathsf{TPR}(\theta) = \mathsf{P}\Big(\widehat{G}(\theta) = 1 \mid \boldsymbol{X} \in \mathcal{G}_1\Big) = 1 - F_1(\theta)\,, \tag{3}$$

$$\mathsf{FPR}(\theta) = \mathsf{P}\Big(\widehat{G}(\theta) = 1 \mid \boldsymbol{X} \in \mathcal{G}_0\Big) = 1 - F_0(\theta)\,. \tag{4}$$

For each cut-point $\theta \in \mathbb{R}$ we obtain a decision of the classifier that can be displayed as a point in so-called *ROC space.* In the ROC space FPR is plotted on the horizontal coordinate and TPR on the vertical one. The best classification corresponds to the top-left corner. Images of useless classifications (so-called *random* ones) are on the main diagonal. A random classifier predicts an object $\boldsymbol{X}$ to be from $\mathcal{G}_1$ with fixed probability p, where $p \in [0,1]$, regardless to the features of $\boldsymbol{X}$, for more details see [1]. The ROC curve combines FPR and TPR managing them for all possible cut-points $\theta \in \mathbb{R}$ at once, i.e.,

$$\mathsf{ROC} \equiv \Big\{ \big[\mathsf{FPR}(\theta); \mathsf{TPR}(\theta)\big],\, \theta \in \mathbb{R} \Big\}\,. \tag{5}$$

This approach allows to select the best cut-point θ for a given criterion, see [4] or [5]. It is useful to express the ROC curve as a function of the horizontal coordinate of ROC graph

$$\mathsf{ROC}(t) = 1 - F_1\big(F_0^{-1}(1-t)\big), \qquad t \in [0,1]\,. \tag{6}$$

There are several approaches to the problem of the ROC curve estimation. Among the parametric ones, there belongs the well-known *binormal* model assuming normality of the score in $\mathcal{G}_1$ and $\mathcal{G}_0$. In fact, it is more widely applicable thanks to the invariancy of the ROC curves to the increasing transformation of the score function, see [4]. The binormal model can be hence used if there exists such an increasing transformation ϕ of Y that $\phi(Y)$ is normal. For the purpose of classifiers comparison, there is a test of equivalency of ROC curve proposed by Metz and Kronman, see [3]. Nevertheless, for our data—see Section 3—it cannot be used. But it is possible to use another parametric model—the *biexponential* one. In Section 2 we derive its estimators, their distribution and the distribution of the equivalency test statistic. Finally, in Section 3, we apply the model to sociological data.

2. Biexponential model

Supposing an exponential distribution of the scores in both groups, we obtain one of the simplest parametric model. The choice of the single parameter $\lambda = \mathsf{E}\,Y$ determines all the moments of the distribution. This simplicity allows to derive the exact distribution of the parameter estimates. On the other side, it constrains the use of the model.

Let us denote the scores in group $\mathcal{G}_j$, $j = 0, 1$ as Y_j. Assume that Y_0 and Y_1 are independent and *exponentially* distributed, i.e., Y_j has the density

$$f_j(y) = \frac{1}{\lambda_j}\,\mathrm{e}^{-\frac{y}{\lambda_j}}\,, \qquad y > 0, \quad \lambda_j > 0, \quad j = 0, 1\,. \tag{7}$$

Consequently, the ROC curve can be expressed in a functional form as

$$\mathsf{ROC}(t) = 1 - F_1\left(F_0^{-1}(1-t)\right) = \exp\left\{\frac{\lambda_0}{\lambda_1}\log(t)\right\} = t^{\zeta}, \tag{8}$$

where $t \in [0, 1]$ and $\zeta = \frac{\lambda_0}{\lambda_1}$.

If $\lambda_0 = \lambda_1$, the curve (8) is the *random* classifier ROC curve. If $\lambda_0 < \lambda_1$, the ROC curve is concave, lying above the main diagonal, and the corresponding classifier is better than the random one. The last option, $\lambda_0 > \lambda_1$, leads to a convex ROC curve indicating that the labels of groups $\mathcal{G}_0$, $\mathcal{G}_1$ are possibly swapped. All the above mentioned situations are displayed in Figure 1.

2.1. Estimation of the parameter ζ

Let $Y_{j1}, \ldots, Y_{jn_j} \overset{iid}{\sim} \mathsf{Exp}(\lambda_j)$, for $j = 0, 1$. Both the use of the *method of moments* (MoM) and the *maximum likelihood method* (ML) lead to the following estimators of λ_j, i.e.,

$$\widehat{\lambda}_j = \overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ji}\,, \qquad j = 0, 1\,. \tag{9}$$

Thanks to the *plug-in* principle of ML estimates (Zehna), the ML estimator of parameter ζ is

$$\widehat{\zeta} = \frac{\overline{Y}_0}{\overline{Y}_1} = \frac{n_1}{n_0}\frac{U_0}{U_1}\,, \qquad \text{where} \quad U_j = \sum_{i=1}^{n_j} Y_{ji}\,, \qquad j = 0, 1\,. \tag{10}$$

We will further deal with the random variable $B = \frac{\widehat{\zeta}}{\zeta}$, because the distribution of $\widehat{\zeta}$ depends on ζ. The following theorem summarizes the properties of B.

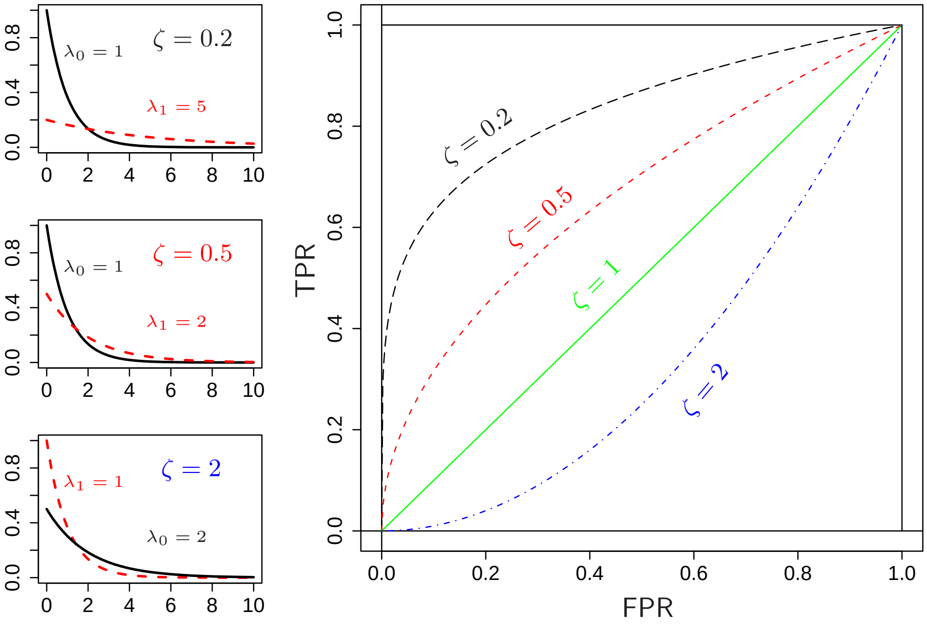

FIGURE 1. Dependence of the ROC curve on intensities λ_0 and λ_1. In the left column of the figure there are densities for different choices of λ_0 and λ_1. The solid lines correspond to the group $\mathcal{G}_0$, while the dashed lines to the group $\mathcal{G}_1$. The relevant ROC curves are shown in the right of the figure.

THEOREM 1. *The random variable* $B = \frac{\widehat{\zeta}}{\zeta}$ *has the density*

$$f_B(b) = \frac{b^{n_0-1}}{\beta(n_0, n_1)} \left(\frac{n_0}{n_1}\right)^{n_0} \left(1 + \frac{n_0}{n_1} b\right)^{-(n_0+n_1)}, \qquad b > 0\,. \tag{11}$$

Proof. Using the expression (10), the random variable B can be rewritten as

$$B = \frac{\widehat{\zeta}}{\zeta} = \frac{n_1}{n_0} \frac{U_0}{\lambda_0} \frac{\lambda_1}{U_1}\,. \tag{12}$$

From the well-known properties of the exponential distribution it follows that $U_j \sim \mathsf{Gamma}\left(\frac{1}{\lambda_j}, n_j\right)$, $j = 0, 1$. The statistics U_0 a U_1 are independent, thus it is possible to derive the density of the variable $Q = \frac{U_0}{U_1}$ by means of the *density-transformation theorem.*

There is an alternative approach using the properties of the Gamma distribution. It is easy to show that if $G \sim \mathsf{Gamma}(a,p)$, for $a,p>0$, then $aG \sim \mathsf{Gamma}(1,p)$. Hence, the random variable $\frac{2U_j}{\lambda_j}$ has the distribution $\mathsf{Gamma}\big(\frac{1}{2},n_j\big)$ which is the chi-square distribution with $2n_j$ degrees of freedom. Consequently, the statistic B has the *Fisher-Snedecor* distribution with $2n_0$ and $2n_1$ degrees of freedom, i.e., the density given by (11). □

Remark. The knowledge of the distribution of B can be used for a one-sample test of the hypothesis $H_0 : \zeta \geq \zeta_0$ vs. $H_1 : \zeta < \zeta_0$. Notice that the choice of $\zeta_0 = 1$ enables comparison of any classifier with the random one, see Figure 1.

2.2. Test of the equivalence of two classifiers in unpaired design

If we compare a pair of classifiers, say A and B, we can use the corresponding ROC curves not only for visualization of their behaviour, but also for the testing whether there is any significant difference between them.

Let us assume that the scores Y^A, Y^B of both classifiers fulfill (at least after some increasing transformation) the *biexponential* model (7), i.e.,

$$Y^C_{j1},\dots,Y^C_{jn^C_j} \overset{iid}{\sim} \mathsf{Exp}\big(\lambda^C_j\big)\,, \qquad \text{for} \quad j=0,1\,, \quad \text{and} \quad C \in \{A,B\}\,. \tag{13}$$

The null hypothesis of the classifiers equivalence can be expressed as $H_0 : \zeta_A = \zeta_B$. Let the experiment be *unpaired*, i.e., for $j=0,1$ the scores $Y^A_{j1},\dots,Y^A_{jn^A_j}$ and $Y^B_{j1},\dots,Y^B_{jn^B_j}$ are independent.

After we establish the notation, we can formulate the main theorem, which allows us to test the hypothesis of equivalence H_0. Put

$$\begin{aligned} T &= \frac{B_A}{B_B} = \frac{\frac{\widehat{\zeta}_A}{\zeta_A}}{\frac{\widehat{\zeta}_B}{\zeta_B}}\,, \\ R &= \frac{n^A_0 n^B_1}{n^A_1 n^B_0}\,, \\ \beta^* &= \frac{\beta\big(n^A_0+n^B_0, n^A_1+n^B_1\big)}{\beta\big(n^A_0,n^A_1\big)\beta\big(n^B_0,n^B_1\big)}\,, \end{aligned} \tag{14}$$

where $\beta(.,.)$ is a beta function.

For $b>0$, $c>0$ and $|x|<1$ let ${}_2F_1(a,b;c;x)$ denotes a Gaussian hypergeometric function, for more details see [2].

Theorem 2. *Under H_0, the statistic T has the density*

$$f_T(t) = \beta^* R^{n^A_0} t^{n^A_0-1}\,{}_2F_1\big(n^A_0+n^A_1, n^A_0+n^B_0; N; 1-Rt\big)\,, \quad 0<t<\frac{2}{R}\,, \tag{15}$$

where

$$N = n_0^A + n_1^A + n_0^B + n_1^B .$$

Especially, if

$$n_0^A = n_0^B = n_0 \quad \textit{and} \quad n_1^A = n_1^B = n_1 ,$$

then

$$f_T(t) = \frac{\beta(2n_0, 2n_1)}{\beta(n_0, n_1)^2} t^{n_0 - 1} \cdot {}_2F_1\big(n_0 + n_1, 2n_0; \, 2(n_0 + n_1); \, 1 - t\big), \tag{16}$$

for $0 < t < 2$.

Proof. Thanks to the independence of A and B, the joint density of $\boldsymbol{B} = (B_A, B_B)^T$ is the product of marginal densities of B_A and B_B given by (11). We define the transformation $t: \ (B_A, B_B)^T \to (B_A/B_B, B_B)^T = (T, Y)^T$ and its inverse $\tau: \ (T, Y)^T \to (TY, Y)^T = (B_A, B_B)^T$. It holds $\boldsymbol{B} \in \mathbb{R}_+ \times \mathbb{R}_+$ from where $T > 0$ and $Y > 0$. In addition

$$\det\big(\tau'(t, y)\big) = \left\| \begin{matrix} y & t \\ 0 & 1 \end{matrix} \right\| = y \, .$$

For $\boldsymbol{t} = (t, y)^T \in \mathbb{R}_+ \times \mathbb{R}_+$, the random vector $\boldsymbol{T} = (T, Y)^T$ has the density

$$f_{\boldsymbol{T}}(\boldsymbol{t}) = y \cdot f_{\boldsymbol{B}}\big(\tau(\boldsymbol{t})\big) = y \cdot f_{B_A}(yt) \cdot f_{B_B}(y) \, , \tag{17}$$

so that for $t > 0$ the density of T is of the form

$$f_T(t) = \frac{K_A K_B t^{n_0^A - 1}}{\beta\big(n_0^A, n_1^A\big)\beta\big(n_0^B, n_1^B\big)} \int_0^{+\infty} \frac{y^{b-1}}{(1 + K_A t y)^{L_A} (1 + K_B y)^{L_B}} \, \mathrm{d}y \, , \tag{18}$$

where

$$b = n_0^A + n_0^B ,$$

$$L_A = n_0^A + n_1^A , \qquad L_B = n_0^B + n_1^B ,$$

$$K_A = \frac{n_0^A}{n_1^A} , \qquad K_B = \frac{n_0^B}{n_1^B} \, .$$

The integral in (18) can be easily transformed into the form

$$\int_0^1 \frac{v^{b-1}(1 - v)^{L_A + L_B - b - 1}}{\big(1 - v(1 - Rt)\big)^{L_A}} \, \mathrm{d}v \tag{19}$$

using the transformations $y \to z = K_B y \to v = \frac{z}{(1+z)}$. Finally, we use the *Euler's integral representation* of the hypergeometric function ${}_2F_1(a, b; c; x)$

$${}_2F_1(a, b; c; x) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c - b)} \int_0^1 \frac{v^{b-1}(1 - v)^{c-b-1}}{(1 - xv)^a} \, \mathrm{d}v \, , \tag{20}$$

which is valid for $b > 0$, $c > 0$ and $|x| < 1$. □

Remark. The representation (15) of the density f_T based on the hypergeometric function ${}_2F_1(a, b; c; x)$ diverges for $t \geq \frac{2}{R}$. This constraint can be overcome by the use of slightly modified version of the test statistic T, i.e., using

$$T^* = \frac{\min(B_A, B_B)}{\max(B_A, B_B)}\,, \tag{21}$$

so that $0 < T^* \leq 1$ and we are concerned with the lower part of critical region.

3. Application to the data

Social scientists of the Department of Sociology of the Faculty of Arts and Philosophy of the Charles University in Prague performed a large survey called Akter in 2005. One of the problems they dealt with was the prediction of the respondents' anxiety of the security situation in their neighborhood by means of some set of predictors that can be measured more easily, especially in smaller surveys. `Anxiety` is treated as binary variable `Anx` with asymmetric prevalence of its modalities $\big(\mathsf{P}(\texttt{Anx} = 1) = 0.32\big)$. The predictors, selected through Principal Component Analysis, are variables such as `Sex`, `Education`, `Trust`, etc.

We used this data set to compare two classic classifiers—*Linear Discriminant Analysis* (LDA) and *Support Vector Machines* (SVM). The data set contains 1849 observations that were randomly split into the *training* set (1387 obs.) and the *test* set (462 obs.), such that the prevalence of anxiety was preserved in both sets.

Consequently, to keep the classifiers independent, the test set was divided into two independent samples once again, preserving the same ratio $\frac{(\texttt{Anx}=1)}{(\texttt{Anx}=0)}$. Then, the score Y_{LDA} was transformed into $Y^*_{LDA} = Y^2_{LDA}$ to have the exponential distribution. and computational simplicity.

The ROC curves of the classifiers are in Figure 2. They seem to be very close. However, the LDA looks slightly better. The question is whether there is any significant difference between them. In the right part of Figure 2, there is plotted the distribution of the test statistic T computed by the statistical package `R`, using the function `hypergeo` from the library `Davies`. The performance of the function is much better for $t < 1$ than for $t > 1$, which is good for the test statistic T^*. Its value is $T = 0.9379$, which is much larger than the 2.5%-quantile (in accordance, the p-value is 0.7483), hence the test does not reject the null hypothesis on the 5% significance level. Neither of the classifiers outperforms significantly the other, thus we can use the LDA because of its comprehensibility to non-mathematicians

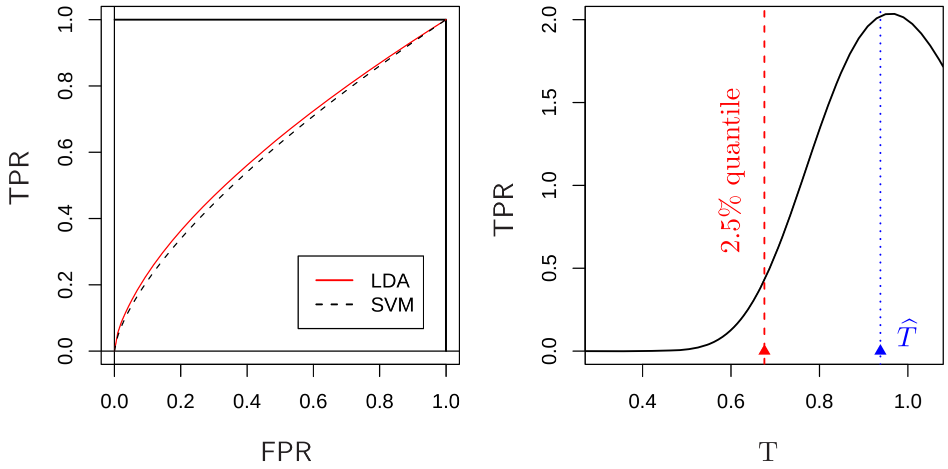

Figure 2. Result of the test for Akter data. ROC curves of the classifiers are in the left subfigure. Distribution of the test statistic T is on the right. The observed value of $T = 0.9379$ is denoted as $\widehat{T}$.

4. Discussion

This paper has presented a parametric model for the ROC curve of a classifier with exponentially distributed scores. This model can be applied in the situations when the other well-known parametric models (e.g., the *binormal* one) cannot be used.

To distinguish between a pair of classifiers, the test of equivalency of two ROC curves has been developed and applied to sociological data.

REFERENCES

[1] FAWCETT, T.: *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*. Hewlett-Packard Laboratories Palo Alto, 2003.

[2] JOHNSON, N. L.—KOTZ, S.—KEMP, A. W.: *Univariate Discrete Distributions*, (2nd ed.), John Wiley & Sons, New York, 1993.

[3] METZ, CH. E.—KRONMAN, H. B.: *Statistical significance tests for binormal ROC curves*. Journal of Mathematical Psychology **22** (1980), 218–243.

[4] PEPE, M. S.: *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York, 2003.

[5] ZHOU, X. H.—OBUCHOVSKI, N. A.—McCLISH, D. K.: *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, New York, 2002.

Received October 2, 2006

Department of Sociology
Faculty of Philosophy and Arts
Charles University Prague
Celetná 20
116 42 Prague 1
CZECH REPUBLIC
E-mail: betinec@matfyz.cz