ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 196, 2006

*Agnieszka Rossa**

CLASSIFICATION TREE BASED ON RECEIVER OPERATING CHARACTERISTIC CURVES

Abstract. The paper deals with a new classification algorithm for discriminating between two populations. The proposed algorithm uses properties of a receiver operating characteristic function $ROC(v)$ and a goodness-of-fit statistic proposed for testing the null hypothesis $H_0: ROC(v) = v$ against $H_1: \sim H_0$.

Key words: classification tree, Receiver Operating Characteristic curve, goodness-of-fit test.

1. THE MAIN IDEA AND NOTATION

Consider the problem of classifying individuals into one of two populations π_0 or π_1. We assume that values of s continuous random variables $X_1, X_2, ..., X_s$ are observed. Variables $X_1, X_2, ..., X_s$ will be called – diagnostic variables.

Let as assume that an individual is to be classified to the population π_0 if X_j exceeds a threshold x_{0_j} for some $j = 1, 2, ..., s$. Assume the following notation

$$Z_j = X_j|\pi_1, \quad j = 1, 2, ..., s, \tag{1}$$

$$C_j = X_j|\pi_0, \quad j = 1, 2, ..., s, \tag{2}$$

and

$$\mathbf{X}^T = [X_1, X_2, ..., X_s, Y], \tag{3}$$

$$\mathbf{Z}^T = [Z_1, Z_2, ..., Z_s, Y], \tag{4}$$

$$\mathbf{C}^T = [C_1, C_2, ..., C_s, Y], \tag{5}$$

* Ph.D., Chair of Statistical Methods, University of Łódź.

where Y is a binary (response) variable indicating whether an observation $[X_1, X_2, ..., X_s]$ comes from π_0 or π_1. Thus,

$$Y = \begin{cases} 0, & \text{if an observation } [X_1, X_2, ..., X_s] \text{ comes from } \pi_0, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

It follows that $Y \equiv 1$ in (4) and $Y \equiv 0$ in (5).

The cumulative distribution function (CDF) of X_j in the populations π_0 and π_1 will be denoted by F_j and G_j, respectively. Using the notation (1)–(2), F_j is a CDF of a random variable C_j and G_j is a CDF of a random variable Z_j.

Assume that we have the learning data which comprise a sequence of n independent copies of the vector (3) given $\mathbf{X}$ comes from the population π_1, and another sequence of m independent copies of (3) given $\mathbf{X}$ comes from π_0 Using the notation (4)–(5), both sequences (random samples) will be denoted in the following form

$$\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_n, \tag{7}$$

and

$$\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_m. \tag{8}$$

The proposed classification algorithm uses properties of a receiver operating characteristic function $ROC(v)$ and a goodness-of-fit statistic used for testing the hypothesis $H_0: ROC(v) = v$ against $H_1: \sim H_0$.

The ROC curve and some of its properties are studied in Section 2, the proposed goodness-of-fit statistic is described in Section 3. The classification procedure is presented in Section 4.

2. THE RECEIVER OPERATING CHARACTERISTIC CURVE

For simplicity, let us assume one diagnostic variable X, with a CDF F if X comes from π_0 or with a CDF G if X comes from π_1. Using the notation (1)–(2) F is a CDF of a random variable C and G is a CDF of a random variable Z.

The receiver operating characteristic curve is a plot of $1 - F(x)$ against $1 - G(x)$ as x varies over the support of X. In other words, it is a plot of $P(X > x | \pi_0)$ against $P(X > x | \pi_1)$ or a plot of $P(C > x)$ against $P(Z > x)$ as the threshold x varies. The ROC curve can be also defined as a set of points of the form

$$\{(1-G(x), 1-F(x)): \quad x \in (-\infty, \infty)\}$$

In statistical terms, the *ROC* curve displays the trade-off between power and size of a test with a rejection region $P(X > x)$ as x is varied. In the biomedical context π_0 is often a disease group and π_1 is a control group. The power $P(X > x|\pi_0)$ is then the probability of a true positive diagnosis and the size $P(X > x|\pi_1)$ is the probability of false positive diagnosis (Green, Swets 1966; Thomas, Myers 1972; Lloyd 1998, 2002).
If X is continuous, then *ROC* depends on F, G via the formula

$$ROC(v) = 1 - F(G^{-1}(1-v)), \quad v \in [0, 1]. \tag{9}$$

Indeed, let us denote

$$v = 1 - G(x),$$

then

$$G(x) = 1 - v \quad \text{and} \quad x(v) = G^{-1}(1-v).$$

Thus

$$ROC(v) = 1 - F(x(v)) = 1 - F(G^{-1}(1-v)) \quad \text{and} \quad v \in [0, 1].$$

$ROC(v)$ is always a non-decreasing function on the unit space, as shown in the example 1 (cf. Figure 3). Estimation of $ROC(v)$ is usually based on replacing F and G by their empirical versions F_m and G_n defined as follows

$$G_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i \leqslant x), \tag{10}$$

$$F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(C_i \leqslant x), \tag{11}$$

where $\mathbf{1}(A)$ denotes a characteristic function of an event A.

The empirical *ROC* curve will be denoted by $R\hat{O}C$. It is a plot of $1 - F_m(x)$ against $1 - G_n(x)$. In other words, an empirical *ROC* curve is a set of points

$$\{(1-G_n(x), 1-F_m(x)): \quad x \in (-\infty, \infty)\}.$$

Example 1

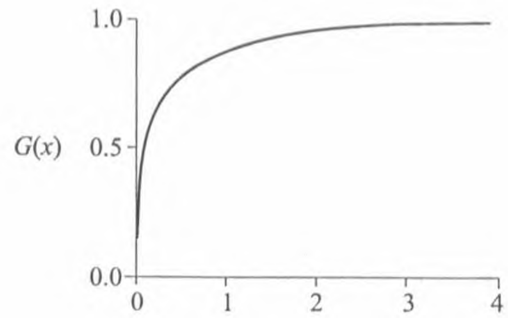

Fig. 1. Cumulative distribution function $G(x)$ of X in π_1

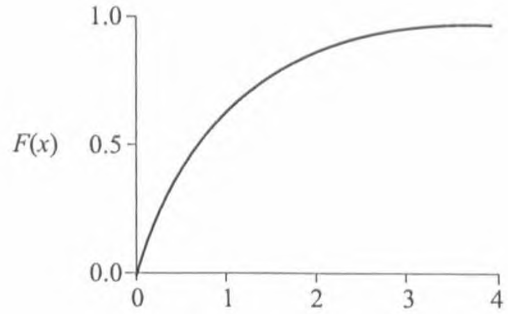

Fig. 2. Cumulative distribution function $F(x)$ of X in π_0

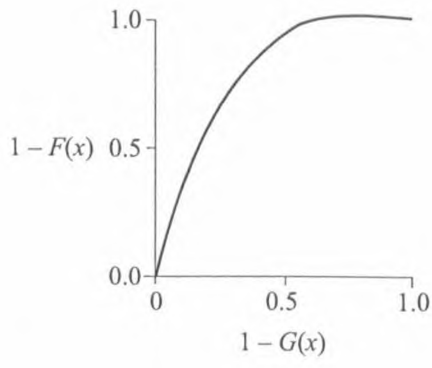

Fig. 3. The *ROC* curve

It is easy to check that the area under the ROC curve (AUC) equals to the probability $P(Z < C)$ Let us first calculate AUC. From (9) we have

$$AUC = \int_0^1 ROC(v)dv = \int_0^1 [1 - F(G^{-1}(1 - v))]dv = \int_0^1 dv - \int_0^1 F(G^{-1}(1 - v))dv =$$

$$= 1 + \int_1^0 F(G^{-1}(v))dv = 1 - \int_0^1 F(G^{-1}(v))dv = 1 - \int_{-\infty}^{\infty} F(c)dG(c). \tag{12}$$

Now, we find the probability $P(Z < C)$. Denote by $f_{(Z,C)}(z, c)$ the two-dimensional density function of (Z, C) and by g, f – the marginal density functions of Z, C, respectively. From independence of Z and C we get $f_{(Z,C)}(z, c) = g(z)f(c)$. Thus,

$$P(Z < C) = P_{(Z,C)}\{(z, c) : z < c\} = \iint_{\{(z,c):z<c\}} f_{(Z,C)}(z, c)dzdc =$$

$$= \iint_{\{(z,c):z<c\}} g(z)f(c)dzdc = \int_{-\infty}^{\infty} f(c)\left[\int_{-\infty}^{c} g(z)dz\right]dc =$$

$$= \int_{-\infty}^{\infty} f(c)G(c)dc = \int_{-\infty}^{\infty} G(c)dF(c) = 1 - \int_{-\infty}^{\infty} F(c)dG(c). \tag{13}$$

Comparing the results (12) and (13) we receive the equality

$$AUC = P(Z < C). \tag{14}$$

It follows from (14) that the ROC curve summarizes the separation between two distributions F and G. The higher is the ROC curve, the greater the prediction accuracy of the diagnostic variable X. If the plot of $ROC(v)$ lies on the diagonal $y = v$ than there are no difference in distributions of the populations π_0 and π_1. In the case of an empirical ROC curve we may state that the more significant is the difference between the empirical ROC curve and a diagonal line on the interval $[0, 1]$, the more significant is the corresponding diagnostic variable X with respect to its prediction accuracy. This concept constitutes the background for the χ^2 goodness-of-fit test discussed in details in the next section.

3. THE GOODNESS-OF-FIT TEST FOR ROC

Consider the null hypothesis of the form

$$H_0 : \forall_{v \in [0, 1]} \; ROC(v) = v, \tag{15}$$

against the alternative

$$H_1 : \sim H_0.$$

Let us notice that $ROC(v)$ defined in (9) is a CDF of the random variable

$$W = 1 - G(C), \tag{16}$$

for we have

$$P(W < v) = P(1 - G(C) < v) = P(G(C) > 1 - v) = P(C > G^{-1}(1 - v)) =$$
$$= 1 - P(C \leqslant G^{-1}(1 - v)) = 1 - F(G^{-1}(1 - v)) = ROC(v), \quad v \in [0, 1]$$

It follows, that testing (15) can be reduced to the problem of testing the hypothesis that W (or equivalently $G(C)$) has the uniform distribution on the unit interval. Hence, the null hypothesis (15) can be reformulated equivalently as

$$H_0 : G(C) \sim \text{Uniform on } [0, 1]. \tag{17}$$

Unfortunately, in order to test (17) we would need to observe a random variable $G(C)$, what is usually impossible without any parametric assumptions concerning the cumulative distribution function G. For this reason we will consider the empirical cumulative function G_n defined in (10) instead of G.

It is easy to notice, that the random variable $G_n(C)$ has a discrete distribution, for it takes the values from the finite set

$$\left\{0, \frac{1}{n}, \frac{1}{n}, \dots, \frac{n-1}{n}, 1\right\}.$$

We will find the probability distribution of $G_n(C)$. Let R, F be $CDFs$ of $G(C)$ and C, respectively. Let us also denote by r and f respective density functions of $G(C)$ and C. For any $i \in \{0, 1, \dots, n\}$ we have

$$P\left(G_n(C) = \frac{i}{n}\right) = \binom{n}{i} \int_{-\infty}^{\infty} G^i(x)[1 - G(x)]^{n-i} f(x)dx.$$

Denote

$$y = G(x),$$

then

$$x = G^{-1}(y), \quad dx = [G^{-1}(y)]' dy.$$

Notice that

$$R(x) = F(G^{-1}(x)) \quad \text{for} \quad x \in [0, 1]$$

and

$$r(x) = f(G^{-1}(x))[G^{-1}(x)]'.$$

Hence

$$P\left(G_n(C) = \frac{i}{n}\right) = \binom{n}{i}\int_0^1 y^i(1-y)^{n-i}f(G^{-1}(y))[G^{-1}(y)]'dy =$$

$$= \binom{n}{i}\int_0^1 y^i(1-y)^{n-i}r(y)dy. \tag{18}$$

If the hypothesis (17) is true then $r(x) = 1$ for $x \in [0, 1]$ and from (18) we have

$$P\left(G_n(C) = \frac{i}{n}|H_0\right) = \binom{n}{i}\int_0^1 y^i(1-y)^{n-i}dy = \frac{1}{n+1}. \tag{19}$$

Assume that we observe a random sample

$$G_n(C_1),\ G_n(C_2),\ \dots,\ G_n(C_m). \tag{20}$$

Now we can use the standard χ^2 goodness-of-fit test for testing (17) with the χ^2 statistic of the form

$$\chi^2 = \sum_{i=0}^{n} \frac{(m_i - mp_i)^2}{mp_i}, \tag{21}$$

where m is the size of the sample (20), p_i represents the hypothetical probability (19) that $G_n(C) = i/n$, and m_i stands for the empirical number of observations in (20) equal to i/n.

It is well known that the statistic (21) under H_0 has an asymptotic χ^2 distribution with n degrees of freedom. Thus, if the sample size m is large, we can use this statistic to test the null hypothesis (17).

4. CLASSIFICATION ALGORITHM

Using the properties of *ROC* curves and the goodness-of-fit statistic discussed in previous sections we will now describe a simple classification rule based on a continuous diagnostic variable. The rule is as follows.

From the set $X_1, X_2, \dots, X_s$ choose a variable X_k, say, for which the goodness-of-fit statistic χ^2 defined in (21) is the largest one. Construct the corresponding empirical curve $\hat{ROC}_k$ and find such a point $x = x_0$ for which the distance between points $(1 - G_n(x), 1 - F_m(x))$ and $(0, 1)$ is the smallest one. The threshold x_0 can be treated as the most predictive one. Suppose that we observe a realization x_k of the variable X_k coming from one of the populations π_0 or π_1. We will classify this observation to π_0 if $x_k > x_0$ and to π_1, otherwise.

Now we can formulate a more complex partitioning procedure employing the whole set of continuous diagnostic variables $X_1, X_2, ..., X_s$. This procedure will be called a learning procedure for it uses the learning sample (7)–(8). It leads to a classification tree that can be used to classify new individuals:

1. Determine the set $\mathcal{N}^{\{l\}}$ of individuals constituting the sample under analysis in the l-th step of the procedure. In the first step the set $\mathcal{N}^{\{1\}}$ consist of all the individuals of the learning sample.

2. For each X_j calculate the χ^2 goodness-of-fit statistic (21). In calculations use observations of X_j for those individuals which belong to $\mathcal{N}^{\{l\}}$.

3. Choose the diagnostic variable X_k for which the χ^2 statistic is the largest one.

4. For the $\hat{ROC}_k$ curve corresponding to X_k find the most predictive threshold x_{0k}.

5. If the realization x_k of X_k for an individual from $\mathcal{N}^{\{l\}}$ is greater than x_{0k}, classify it to π_0, otherwise – to π_1. Repeat the step for all the individuals in $\mathcal{N}^{\{l\}}$.

6. Denote by $\mathcal{N}_0^{\{l\}}$ the set of individuals from $\mathcal{N}^{\{l\}}$ classified to π_0 and by $\mathcal{N}_1^{\{l\}}$ the set of individuals classified to π_1. If for all the individuals in $\mathcal{N}_0^{\{l\}}$ the variable Y defined in (6) equals to 0 then treat the set $\mathcal{N}_0^{\{l\}}$ as a terminal one. If all the individuals in $\mathcal{N}_1^{\{l\}}$ have $Y = 1$ then treat $\mathcal{N}_1^{\{l\}}$ also as terminal set. If one of (or both) sets $\mathcal{N}_0^{\{l\}}$ and $\mathcal{N}_1^{\{l\}}$ are non-homogenous with respect to Y than take the given non-homogenous set as $\mathcal{N}^{\{l+1\}}$ and return to the first step of the procedure.

The procedure continues until the resulting sets contain individuals homogenous with respect to Y.

REFERENCES

Green D. M., Swets J. A. (1966), *Signal Detection Theory and Psychophysics*, Wiley, New York.

Lloyd C. J. (1998), "Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems", *Journal of American Statistical Association*, **93**, 1356–1364.

Lloyd C. J. (2002), "Semi-Parametric Estimation of ROC Curves Based On Binomial Regression Modelling", *Australian and New Zeland Journal of Statistical*, **44**, 75–86.

Thomas A. N., Myers J. L. (1972), "Implications of Latency Data for Threshold and Non-Threshold Models of Signal Detection", *Journal of Mathematical Psychology*, **9**, 253–285.

Agnieszka Rossa

DRZEWO KLASYFIKACYJNE OPARTE NA KRZYWYCH OPERACYJNO-CHARAKTERYSTYCZNYCH

(Streszczenie)

W artykule przedstawiono propozycję konstrukcji drzewa klasyfikacyjnego, wykorzystującą własności krzywych operacyjno-charakterystycznych oraz statystyki testu zgodności χ^2 dla weryfikacji hipotezy zerowej $H_0: ROC(v) = v$ przeciwko hipotezie $H_1: \sim H_0$.