10.2478/v10048-009-0023-5

Receiver Operating Characteristic Analysis for Classification Based on Various Prior Probabilities of Groups with an Application to Breath Analysis

K. Cimermanová

Institute of Measurement Science, Department of Theoretical Methods, Slovak Academy of Sciences,
Dúbravská cesta, 9, 841 04, Bratislava, Slovakia, katarina.cimermanova@gmail.com

In this paper we illustrate the influence of prior probabilities of diseases on diagnostic reasoning. For various prior probabilities of classified groups characterized by volatile organic compounds of breath profile, smokers and non-smokers, we constructed the ROC curve and the Youden index with related asymptotic pointwise confidence intervals.

Keywords : Breath analysis, ROC analysis, confidence intervals, discriminant analysis, prior probability

1. INTRODUCTION

THE PRIOR PROBABILITIES are independent of measured data and known before taking any observations. The change of this information changes the probability of a correct output of a diagnostic test [5]. In this paper we show how radical these changes are in classification of measured concentrations of volatile organic compounds of smokers and non-smokers.

The ROC (receiver operating characteristic) curve is a metric for comparing predicted and actual target values in a classification model. The ROC curve plots sensitivity and (1 - specificity) of the diagnostic test. The sensitivity measures the proportion of actual positives which are correctly identified as such (i.e. the percentage of sick people who are identified as having the condition); and the specificity measures the proportion of negatives which are correctly identified (i.e. the percentage of healthy people who are identified as not having the condition).

Different classification algorithms use different techniques for finding relationships between the measured values of subjects (e.g. concentrations of selected volatile organic compounds, VOCs, of breath profile) and the known targets (association with groups, e.g. smokers or non-smokers). We use the discriminant function $g(\boldsymbol{X})$ with a threshold (the decision point used by the model for classification) dependent on prior probabilities of groups, [5]. The ROC curve measures the impact of changes in the threshold. For the ROC curve related to changes of prior probabilities we constructed the asymptotic pointwise confidence interval, [4] (CI describes the range where the true ROC curve lies with some specific probability, e.g. 95% CI).

To evaluate effectiveness of classification based on different prior probabilities of discriminated classes we use the Youden index [3]. This index ranges between 0 and 1, with a value close to 1 indicating that the effectiveness of algorithm is relatively large and a value close to 0 indicating limited effectiveness. For the Youden index we constructed the asymptotic pointwise confidence interval, too.

We apply the classification on breath analysis data. Breath analysis as a non-invasive technique is very attractive because it can be easily applied to sick patients, including children and elderly people. It offers potential for detection of some diseases, e.g. diabetes, lung and esophageal cancer etc. In our study we consider measured values of breath profile of smokers and non-smokers measured by proton transfer reaction mass spectrometry PTR-MS, for more details see e.g. [1]. The molecular masses detectable by the PTR-MS range from m/z 21 to m/z 230. The selected compounds (m/z values) for our analysis are m/z 28 (tentatively identified as hydrogen cyanide), m/z 31 (formaldehyde), m/z 42 (acetonitrile), m/z 53 (1-buten-3-alkyne), m/z 59 (acetone), m/z 79 (benzene), m/z 93 (toluene) and further m/z 97, m/z 105, m/z 109 and m/z 123, for more details see [6]. The measured quantities (counts) are transformed [6] to concentrations of volatile organic compounds in ppb (particles per billion) levels. From previous studies [6] we see that the measured data have better properties after logarithmic transformation. Therefore the raw data were log-transformed.

2. ROC ANALYSIS

Let us have a random vector $\boldsymbol{X} = (X_1,\ldots,X_n)$ where X_j represents random variable of log-transformed concentrations of the j-th volatile organic compound (VOC) and n is the number of selected VOCs. For each subject i, $i = 1,\ldots,N$ where N is number of all subjects, defined by measured values $\boldsymbol{x}_i = (x_{i1},\ldots,x_{in})$ we have categorization to a population y_i, i.e. the target

$$\boldsymbol{x}_i \in \omega_1 \Rightarrow y_i = 1$$
$$\boldsymbol{x}_i \in \omega_2 \Rightarrow y_i = -1.$$

For the population of the group of smokers ω_1, and the group of non-smokers ω_2 we assume n-dimensional normal distribution.

For classification we use the quadratic discriminant function

$$\begin{aligned} g(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \\ &+ \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1| \end{aligned} \quad , \qquad (1)$$

where $\boldsymbol{x}$ is a vector of observed values of a subject, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are mean values estimated from training data and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are covariance matrices estimated from training data. (The database is divided into a training and a testing set in some ratio, e.g. 3:2).

For a new observation from the testing set $\boldsymbol{x} = (x_1,\dots,x_n)$ we evaluate the value of the discriminant function $g(\boldsymbol{x})$. This value is compared with a threshold value k, $-\infty < k < \infty$. In our case the threshold value k is defined as

$$k = \ln \frac{P(\omega_2)}{P(\omega_1)} \tag{2}$$

where $P(\omega_1)$ and $P(\omega_2)$ are prior probabilities of group membership, more in [5]. When $g(\boldsymbol{x}) > k$ the subject is classified to the group of positives

$$g(\boldsymbol{x}) > k \Rightarrow \boldsymbol{x} \in \omega_1$$

and otherwise when

$$g(\boldsymbol{x}) \leq k \Rightarrow \boldsymbol{x} \in \omega_2.$$

From results of classification of testing data we can evaluate sensitivity Se and specificity Sp as

$$Se = \frac{TP}{TP + FN} \text{ and } Sp = \frac{TN}{TN + FP} \tag{3}$$

where TP is the number of true positives (positive subject is classified as positive, $\#i$: $g(\boldsymbol{x}_i) > k$ and $y_i = 1$), TN of true negatives (negative subject as negative, $\#i$: $g(\boldsymbol{x}_i) \leq k$ and $y_i = -1$), FP of false positives (negative as positive, $\#i$: $g(\boldsymbol{x}_i) > k$ and $y_i = -1$) and FN is the number of false negatives (positive subject as negative, $\#i$: $g(\boldsymbol{x}_i) \leq k$ and $y_i = 1$).

The sensitivity can be expressed as $Se(k) = P(g(\boldsymbol{X}) > k \mid y = 1) = 1 - P(g(\boldsymbol{X}) \leq k \mid y = 1) = 1 - G(k)$ and specificity $Sp(k) = P(g(\boldsymbol{X}) \leq k \mid y = -1) = F(k)$, where $G(k)$ and $F(k)$ can be interpreted as cumulative distribution functions (cdfs) of discriminant function $g(\boldsymbol{X})$ for positive group ω_1 and negative group ω_2. The alternative definition of the ROC curve is

$$R(1-t) = 1 - G\{F^{-1}(t)\} \tag{4}$$

For $0 \leq t \leq 1$ where F-1$(t) = \inf\{k: F(k) \geq t\}$ denotes the generalized inverse function of F. However, since empirical cdfs $\hat{F}$ and $\hat{G}$ are discontinuous, the estimate of $R(1 - t)$ might have a very erratic appearance [4]. For this reason, it can be advantageous to use smooth empirical cdfs for calculating the estimator of $R(1 - t)$. From empirical cumulative distribution functions we construct estimates of probability density functions (pdfs) f and g, based on normalized histograms. We smooth the functions $\hat{f}$ and $\hat{g}$, too. Next we assume that f and g are continuous and f/g is bounded on any subinterval (a,b) of $(0,1)$, and $n/m \to \lambda$ (a constant) as $\min(n,m) \to \infty$, where n and m are sample sizes of training sets of the populations. The asymptotic pointwise estimate of a CI for $R(1 - t)$ is defined as

$$\hat{R}(1-t) \pm z(\alpha/2)\hat{\sigma}(t) \tag{5}$$

where $z(\alpha/2)$ is the $\alpha/2$-quantile of the standard normal distribution, α is a chosen level of significance and $\hat{\sigma}$ is estimated standard deviation of the ROC curve defined later. In [4] it is shown that a probability space exists on which one can define two independent Brownian bridges $B_1^{(\mathrm{n})}$, $B_2^{(\mathrm{n})}$ such that

$$\begin{aligned} &\sqrt{n}\hat{G}\{\hat{F}^{-1}(t) - G\{F^{-1}(t)\}\} = \\ &\sqrt{\lambda}B_1^{(n)}(G\{F^{-1}(t)\}) + \frac{g(F^{-1}(t))}{f(F^{-1}(t))}B_2^{(n)}(t) + \\ &+ o(n^{-1/2}(\log n)^2) \end{aligned} \tag{6}$$

For a Brownian bridge $B^{(n)}$ we have $\mathrm{E}(B^{(n)}(t)) = 0$ and $\mathrm{E}(B^{(n)}(t)B^{(n)}(s)) = t\ (1\text{-}s)$, for more details see [2]. It can be shown that $\hat{G}\{\hat{F}^{-1}(t)\} - G\{F^{-1}(t)\}$ is asymptotically normally distributed, as

$$\begin{aligned} &\hat{G}\{\hat{F}^{-1}(t)\} - G\{F^{-1}(t)\} \approx \\ &\hat{G}\{F^{-1}(t)\} - G\{F^{-1}(t)\} - \frac{g\{F^{-1}(t)\}}{f\{F^{-1}(t)\}}[\hat{F}\{F^{-1}(t)\} - t] \end{aligned}, \tag{7}$$

with zero mean and variance

$$\begin{aligned} &\sigma^2(t) = \\ &\frac{1}{n}G\{F^{-1}(t)\}[1 - G\{F^{-1}(t)\}] + \frac{1}{m}\frac{g^2\{F^{-1}(t)\}}{f^2\{F^{-1}(t)\}}t(1-t) \end{aligned}, \tag{8}$$

where after replacing F, G, f and g by the respective estimators $\hat{F}$, $\hat{G}$, $\hat{f}$ and $\hat{g}$ we obtain an estimator of σ^2 for $\hat{R}(1-t)$.

The Youden index is defined as

$$J(t) = Se(k) + Sp(k) - 1 \tag{9}$$

for all possible threshold values k [3]. It is the maximum vertical distance between the ROC curve and the diagonal or the chance line, Fig. 1. The Youden index can be rewritten as $J(t) = Se(k) + Sp(k) - 1 = F(k) - G(k) = R(1 - t) + F(k) - 1$, where $F(k) = t$ is regarded as a constant. So for the Youden index we can write an asymptotic pointwise CI

$$\hat{J}(t) \pm z(\alpha/2)\hat{\sigma}(t), \tag{10}$$

where $\hat{\sigma}$ is estimator σ of the ROC curve $\hat{R}(1-t)$. The optimal choice of prior probabilities of groups used in classification is at the point where the Youden index is maximal.

3. Results

Recent results suggest that breath-concentrations could be expected to be log-normally distributed and that the logarithmic transformation of the data could be profitable, e.g. [6]. We verified the assumption of normal distribution of transformed database by the test of skewness and kurtosis of n-dimensional normal distribution [7]. In the database, we have measured concentrations of selected VOCs for 44 smokers and 173 non-smokers.

The sensitivity Se and specificity Sp was estimated by (3), where TN, TP, FP, FN values were computed as arithmetic means based on 100 times divided database in 3:2 ratio for different k defined by the prior probabilities $P(\omega_1) = 0.001{:}0.001{:}0.999$ and $P(\omega_2) = 1 - P(\omega_1)$.

From proportion Se and Sp empirical cdfs of discriminant function $g(\boldsymbol{X})$ were evaluated for the positive group $\hat{G}$ and the negative group $\hat{F}$. The functions $\hat{f}$ and $\hat{g}$, pdfs of discriminant

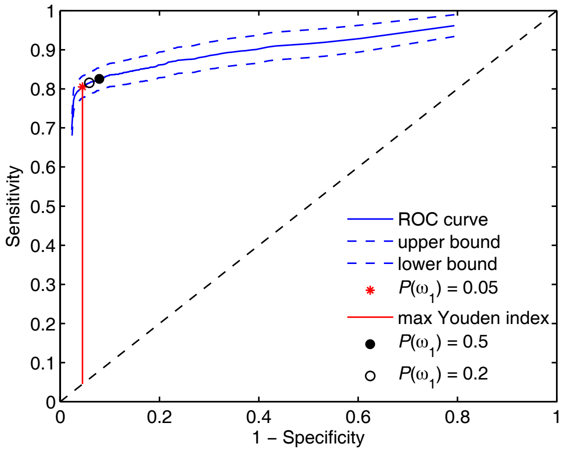

Fig.1 The ROC curve with 95% confidence interval for discriminant function for two groups with threshold dependent on prior probabilities of groups, optimal threshold point with related Youden index and other threshold points characterized by prior probability of positive group.

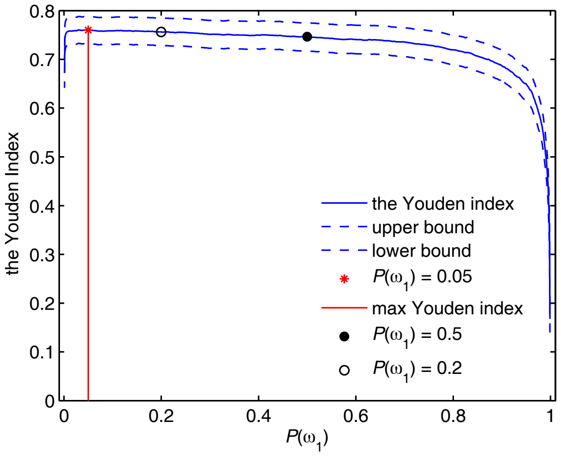

Fig.2 The Youden index with 95% confidence interval for discriminant function for two classes with threshold dependent on prior probabilities of groups and optimal threshold point with related Youden index.

function were computed from normalized histogram from empirical cdfs. For computing the ROC curve by (4) and the standard deviation of the ROC curve by (8), we smoothed $\hat{F}$, $\hat{G}$, $\hat{f}$ and $\hat{g}$ functions with Gaussian window. The Youden index was evaluated by (9) with 95% confidence interval by (10).

The results of classification of smokers and non-smokers are plotted in Fig.1 and Fig.2. The most effective classification is for prior probability of smokers $P(\omega_1) = 0.05$. We also see that the effectiveness of classification is different for different prior probabilities of group membership.

4. Discussion/Conclusions

The ROC analysis is an important tool to summarize the performance (sensitivity and specificity, measures used in medicine) of a medical diagnostic test. By the Youden index we see effectiveness of classification.

The confidence bands are a useful graphical tool for visualizing the statistical variability of the ROC curve and the Youden index estimated from clinical data.

The choice of the prior probabilities of group membership, e.g. insurance company estimates the prior probabilities from knowledge about population but a doctor assesses them based on his own knowledge about the patient, has impact on the results of classification. Our classification method minimizes the total error of classification under the given prior probabilities of group membership. Fig.2 shows the well-known problem: for very high prior probability of one of the groups the classification method has a tendency to classify almost all data to this group. This behavior is not shown in the total error. Therefore we evaluate the Youden index (in place of total error) for the classification of the groups.

Acknowledgment

The research was supported partly by the Slovak Research and Development Agency (APVV), grant SK-AT-0003-08 and RPEU-0008-06, by the Scientific Grant Agency of the Slovak Republic (VEGA), grant No. 2/7087/27 and 1/0077/09.

References

[1] Amann, A., Smith, D. (2005). *Breath Analysis for Clinical Diagnosis and Therapeutic Monitoring.* Singapore: World Scientific.

[2] Billingley, P. (1968). *Convergence of Probability Measure*. New York: John Wiley & Sons.

[3] Fluss, R., Faraggi, D., Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometric Journal*, 47, 45-72.

[4] Hall, P.G., Hyndman, R.J., Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curve. *Biometrika*, 91 (3), 743-750.

[5] Hand, D.J. (1981). *Discrimination and Classification (Wiley Series in Probability and Mathematical Statistics).* John Wiley & Sons.

[6] Kushch, I., et. al. (2008). Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as deremined in a pilot study using PTR-MS. *Journal of Breath Research*, 2, 1-26.

[7] Lamoš, F., Potocký, R. (1998). *Probability and Mathematical Statistic (Statistical Analysis).* Bratislava: Comenius University. (in Slovak)