The ROC Analysis for Classification of Smokers and Non-Smokers Based on Various Prior Probabilities of Groups

K. Cimermanová

Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia
Email: katarina.cimermanova@gmail.com

Abstract. *This paper addresses the influence of the prior probabilities of diseases for diagnostic reasoning. For various prior probabilities of smokers and non-smokers used in discriminant analysis we construct the ROC curve and the Youden Index with related asymptotic pointwise confidence intervals. We show how the prior probabilities change probability of diagnostic results.*

Keywords: Breath Analysis, ROC Analysis, Confidence Intervals, Discriminant Analysis, Prior Probability

1. Introduction

The prior probabilities are independent of measured data and known before we have taken 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 well 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 $g(X)$ 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 threshold. For the ROC curve related with changes of prior probabilities we construct asymptotic pointwise confidence interval, [4] (CI describes 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 construct 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 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 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 levels.

2. ROC analysis

Let us have random vectors $X_i = (x_{i1}, \ldots x_{in})$ where $i = 1, \ldots N$, N is the number of all observed subjects, x_{ij} represents measured concentration of the j-th volatile organic compound (VOC) of subject i and n is the number of selected VOCs. For each subject X_i we have categorization to a population, i.e. the target. For the population of each group (smokers and non-smokers) we assume n-dimensional normal distribution.
For classification we can use the discriminant function

$$g(X) = -\frac{1}{2}(X-\mu_1)'\Sigma_1^{-1}(X-\mu_1) + \frac{1}{2}(X-\mu_2)'\Sigma_2^{-1}(X-\mu_2) + \ln|\Sigma_2|/|\Sigma_1|, \qquad (1)$$

where X is a vector of observed values of a subject, μ_1 and μ_2 are mean values estimated from training data and Σ_1 and Σ_2 are covariance matrices estimated from training data. (The database is divided into training and a testing set in some ratio, e.g. 3:2).

For the new observation X (from the testing set) we evaluate the value of the discriminant function $g(X)$. This value is compared with a threshold value k, $-\infty < k < \infty$. When $g(X) > k$ subject is classify to the group of positives ($X \in \omega_1$) and otherwise when $g(X) < k$, the subject is classify as negative ($X \in \omega_2$). In our case the threshold value k is defined as

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

where $P(\omega_1)$ and $P(\omega_2)$ are prior probabilities, more in [5].

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

$$Se = \frac{TP}{TP+FN} \quad \text{and} \quad Sp = \frac{TN}{TN+FP} \qquad (3)$$

where TP is true positive (positive subject is classified as positive), TN is true negative (negative subject as negative), FP is false positive (negative as positive) and FN is false negative (positive subject as negative).

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

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

for $0 \le t \le 1$ where $F^{1}(t) = \inf\{X\colon F(k) \ge t\}$ denotes the generalized inverse function of F. However, since empirical cdfs are discontinuous, the estimate of $R(1 - k)$ 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 cumulative distribution functions we construct, based on normalized histograms, probability density functions (pdfs) f and g. We smooth the functions f and 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 $n/m \rightarrow \lambda$ as $\min(n,m) \rightarrow \infty$, where n and m are sample size of training sets of populations. The asymptotic pointwise estimate of a CI for $R(1 - k)$ is defined as

$$R(1-t) \pm z(\alpha/2)\sigma(t) \qquad (5)$$

where $z(\alpha/2)$ is the $\alpha/2$-quantile of a standard normal distribution, α is a chosen level of significance and σ is the 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 such that

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

for this processes we have $E(B^{(n)}(t)) = 0$ and $E(B^{(n)}(t)B^{(n)}(s)) = t\,(1-s)$, 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

$$\hat{G}\{\hat{F}^{-1}(t)\} - G\{F^{-1}(t)\} \approx \hat{G}\{F^{-1}(t)\} - G\{F^{-1}(t)\} - \frac{g\{F^{-1}(t)\}}{f\{F^{-1}(t)\}}[\hat{F}\{F^{-1}(t)\} - t] \qquad (7)$$

with zero mean and variance

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

where after replacing F, G, f and g by the respective estimators we obtain an estimator of σ^2 for $R(1 - t)$. The Youden Index is defined as

$$J(t) = Se(k) + Sp(k) - 1 \qquad (9)$$

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), left. The Youden Index can be rewritten as $J(t) = Se(k) + Sp(k) - 1 = F(k) - G(k) = R(1 - k) + F(k) - 1$, where $F(k) = t$ is regarded as a constant. So for the Youden Index we can write an asymptotic pointwise CI

$$J(t) \pm z(\alpha/2)\sigma(t), \qquad (10)$$

where σ is estimated by Eq. (1) defined for σ estimator of the ROC curve $R(1-t)$. The optimal 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]. Therefore we analyze logarithmic transformed data. 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 Eq (3), where TN, TP, FP, FN values were computed as arithmetic means based on 1000 times divided database in ratio 3:2 for different k defined by the prior probabilities $P(\omega_1) = 1{:}0.01{:}1$ and $P(\omega_2) = 1 - P(\omega_1)$.

From proportion Se and Sp ecdfs of discriminant function $g(X)$ were evaluated for the positive group G and the negative group F. The functions f and g, pdfs of discriminant function were computed from normalized histogram from ecdf. For computing of the ROC curve by Eq. (4) and the standard deviation of the ROC curve by Eq. (8), we smoothed G, F, f and g functions with Gaussian window. The Youden Index was evaluated by Eq. (9) with 95% confidence interval by Eq. (10).

The results of classification of smokers and non-smokers are plotted in Fig 1. The most effective classification is for prior probability of smokers $P(\omega_1) = 0.3$. We also see that the effectiveness of classification is different for different prior probabilities.

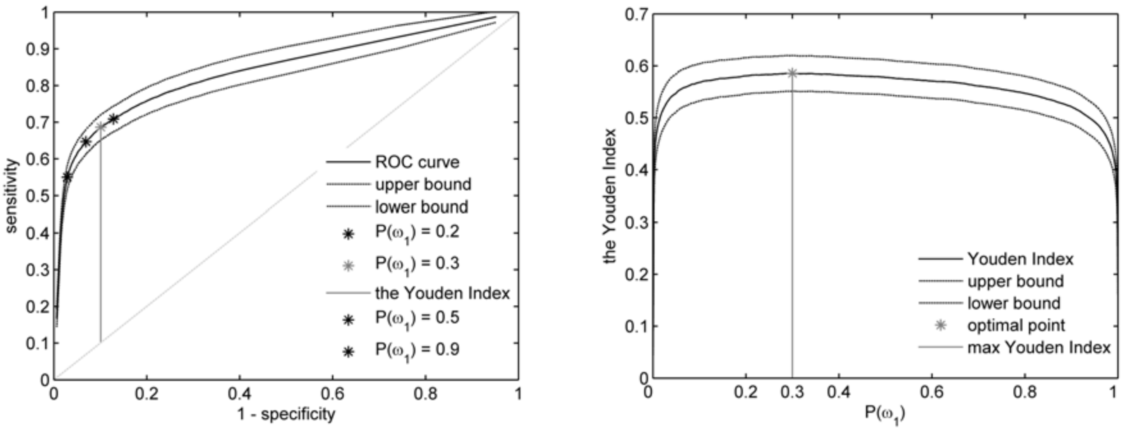

Fig.1. (left) The ROC curve with 95% confidence interval for discriminant function for two groups with threshold dependent on prior probabilities of groups, optimal threshold point with related Youden Index and other threshold points characterized by prior probability of positive group.
(right) The Youden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

4. Discussion and Conclusions

The ROC analysis is an important tool to summarize the performance 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 diagnostic test of clinical data.

Acknowledgements

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 2/7087/27 and 1/0077/09.

References

[1] Amann, A., Smith, D. Breath analysis for clinical diagnosis and therapeutic monitoring, World Scientific, Singapore, 2005

[2] Billingley, P. Convergence of Probability Measure, John Wiley & Sons, Inc., New York, 1968, 64-65

[3] Fluss, R., Faraggi, D., Reiser, B. Estimation of the Youden Index and its Associated Cutoff Point, *Biometric Journal*, Vol. 47, 2005, 45-72

[4] Hall, P.G., Hyndman, R.J., Fan, Y. Nonparametric Confidence Intervals for Receiver Operating Characteristic Curve, *Biometrika*, Vol. 91, No. 3, 2004, 743-750

[5] Hand, D.J. Discrimination and classification, Wiley series in probability and mathematical statistics, J Wiley, 1981

[6] Kushch, I., et al. Compounds Enhanced in a Mass Spectrometric Profile of Smokers' Exhaled Breath Versus Non-smokers as Determined in a Pilot Study Using PTR-MS, *Journal of Breath Research*, Vol.2, 2008, 1-26