The Influence of Increasing Number of Breath Gas Compounds on Binary Classification of Noisy Data

K. Bartošová

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

Abstract. *Classification of subjects into one of two classes is an important problem. Subject's exhaled breath profile measured by Proton Transfer Reaction Mass Spectrometry (PTR-MS) includes 210 different compounds; assigned to mass to charge ratios m/z. Analyzed breath concentrations include variability of repetitive measurements of subjects, thus we label them as noisy data. For classification we use a robust classification method for noisy data.*
In this paper we observe an influence of the number of input compounds on sensitivity and specificity of classification. The order in which the compounds are added is determined by related Youden Indexes.

Keywords: Breath Gas Analysis, Noisy Data, Robust Classification Method, the Youden Index

Introduction

Exhaled breath is a product of inhaled breath and molecules, more specific Volatile Organic Compounds (VOCs), which are released from blood of lung alveolus. Thereto, they are also contributed from cells of upper airways and digestive tract.
In the present time, by development of analytical methods for measuring small quantities of VOCs, we can detect 3481 different compounds in human breath. From them, there are 1753 compounds which have positive alveolar gradient [2]. It means that the concentrations of these compounds are greater in exhaled breath then in inhaled breath. We expect for these compounds that they are endogenous, produced by human body. For this reason the breath analysis is an attractive noninvasive method, without direct intervention to the human body and therefore without a risk for a patient during multiple repetitions.
An ideal analytical method for measurement of small quantities of VOCs in human breath is Proton Transfer Reaction Mass Spectrometry (PTR-MS). PTR-MS is suitable for measuring low concentrations; at particles per billion (ppb) levels. PTR-MS achieves to measure in real time with low detection limit; particles per trillion (ppt) level [1]. The PTR-MS methodology is based on a reaction where the proton H^+ is transferred to VOC from a precursor in drift tube during certain conditions. The precursor, protonated water H_3O^+, is produced by decomposition of water in primary ion source. After the reaction, new ions $VOCH^+$ are selected based on molecular masses, mass to charge ratio m/z, by electromagnetic field and consecutively quantified by ion multiplier into the quadrupol mass spectrometer. Molecular masses detectable by PTR-MS range from m/z 21 to m/z 230. The compounds measured as m/z are tentatively identified as VOC with strongest representation, e.g. m/z 42 is tentatively identified as acetonitrile.
An analyzed data come from a pilot study prepared at Medical University in Innsbruck during the years 2006 and 2008. The breath samples of volunteers were collected to Tedlar bags during a check-up. For some volunteers more bags were collected. Each bag was measured by PTR-MS at least three times.
The database includes data for 217 volunteers, but of which 173 are non-smokers and 44 smokers. For each subject we obtain only one representative vector of exhaled breath profile computed as a median from medians of repetitive measurements of subject bag profiles. Data obtained like this include variability of repetitive measurements. We label them as noisy data.

Let us have a random vector $\boldsymbol{X} = (X_1,\ldots,X_N)$ where X_j represents random variable of concentrations of the j-th compound and N is the number of all compounds. For each subject i, $i = 1,\ldots,n$ where n is the number of all subjects, defined by measured values $\boldsymbol{x}_i = (x_{i1},\ldots,x_{iN})$ we have categorization to a population y_i: y_i=1 if $\boldsymbol{x}_i \in \omega^{(1)}$ and y_i=-1 if $\boldsymbol{x}_i \in \omega^{(2)}$. Superscripts (1), (2) denote affiliation with the positive group $\omega^{(1)}$ and the negative group $\omega^{(2)}$ of subjects respectively.

Selection of Statistically Significant Breath Gas Compounds

Not all measured compounds are statistically significant for classification of subjects based on observed marks. The selection of compounds could be realized based on related Youden indexes. The Youden index measures effectiveness of a compound for distinguishing subjects with observed mark from those without. This index ranges between 0 and 1. A value close to 1 indicates large effectiveness; usually the compound is considered a biomarker of the observed mark. On the other hand a value close to 0 indicates limited effectiveness [4].

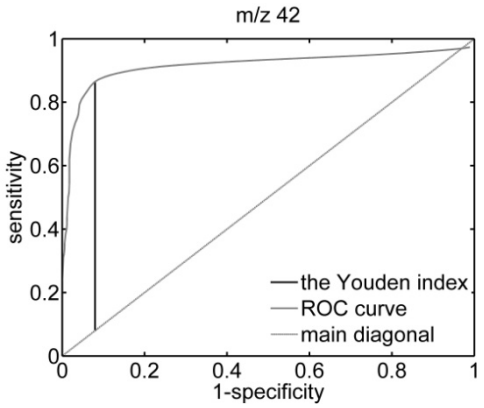

Fig.1. Graphical representation of the Youden index as the maximum vertical distance between the ROC curve and the main diagonal of the graph. The Youden index measures compound effectiveness of classification based on the observed mark.

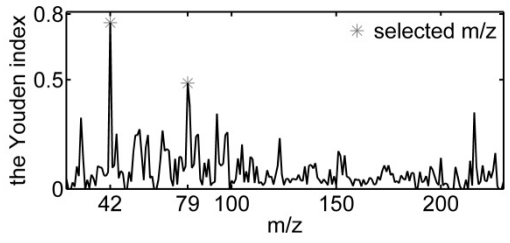

Fig.2. The Youden index for breath compounds measured by PTR-MS with marked selected compounds m/z 42 and m/z 79, tentatively identified as acetonitrile and benzene, for which the Youden index is greater than 0.5.

The Youden index J is a function of sensitivity Se and specificity Sp

$$J = \max_t \{Se(t) + Sp(t) - 1\} \qquad (1)$$

for all points t of the measured variable X_j. Sensitivity Se at a point t is defined as the probability of a correct classification of positive subjects

$$Se(t) = P\left(X^{(1)} > t\right) = 1 - P\left(X^{(1)} \le t\right) = 1 - F^{(1)}(t)$$

and can be evaluated using distribution function $F^{(1)}$ of the positive group at the point t. Specificity Sp at a point t represents the probability of a correct categorization of negative subjects (subjects without the observed mark)

$$Sp(t) = P\left(X^{(2)} \le t\right) = F^{(2)}(t)$$

and corresponds to a value of the distribution function $F^{(2)}$ of the negative group at the point t.

The Youden index J can be rewritten in the form

$$J(t) = Se(t) + Sp(t) - 1 = F^{(2)}(t) - F^{(1)}(t)$$

$$J\left\{F^{(2)-1}(p)\right\} = p - F^{(1)}\left\{F^{(2)-1}(p)\right\},$$

where $F^{(2)-1}(p) = t$ is a value of the p-quantile of the distribution function of the negative group, i.e. $F^{(2)}(t) = p$.

Because empirical distribution functions are not continuous the estimate of the Youden index has a very erratic appearance, especially in the case when the group sizes are different. Therefore to estimate this index we use smoothed estimates of cumulative distribution functions obtained with the help of a Gauss kernel function [5]

$$\hat{F}^{(\cdot)}(t) = \frac{1}{n^{(\cdot)}} \sum_{i=1}^{n^{(\cdot)}} \Phi\left(\frac{t - x_i^{(\cdot)}}{h^{(\cdot)}}\right)$$

where h is the band width of the kernel function and Φ is the distribution function of the standard normal distribution N(0,1).

Classification of Noisy Data

Classification of subjects into one of two classes is an important problem. There are some classification methods which classify data into one of two classes, but in a real life situation the observed vectors are corrupted with noise. A solution to this problem is a robust formulation that stems from the Support Vector Machine (SVM) method. The formulation is a convex optimization problem; in particular, it is an instance of the Second Order Cone Programming (SOCP) problem. An ellipsoidal uncertainty model is assumed; it means that the true value, not always the measured value, is some point of the specified ellipsoid. The classification method assumes only the existence of the second order moments [3].

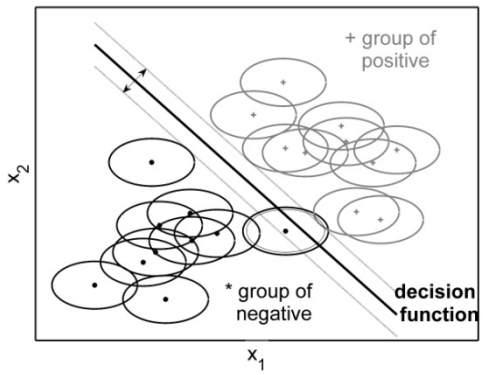

Fig.3. Robust classification method for classification of noisy data, where the optimal linear decision function is found based on specified ellipsoids representing the input data. The optimal decision function is found by the maximization of a margin between two hyperplanes parallel to the decision function with respect to the minimal lost (ellipsoids fallen to other side of the decision function).

Let our classifier be a hyperplane $\langle \mathbf{w}, \boldsymbol{x} \rangle + b = 0$. Our goal is to find optimal parameters $\mathbf{w}$,b based on specified input ellipsoids $\boldsymbol{x} \in B(\boldsymbol{x}_i, \boldsymbol{\Sigma}_i, \gamma)$, where $\boldsymbol{x}_i$ and $\boldsymbol{\Sigma}_i$ are the center and the shape matrix of the i-th input ellipsoids, $i = 1,\ldots,n$, where n is the number of all subjects and γ is a noise level $\gamma \geq 0$. When $\gamma = 0$ we assume input data without noise.
This leads to an optimization problem

$$\min_{\mathbf{w},b,\xi} \sum_{i=1}^{n} \xi_i$$

w.c $$y_i\left(\langle \mathbf{w}, \boldsymbol{x}_i \rangle + b\right) \geq 1 - \xi_i + \gamma_i \left\| \boldsymbol{\Sigma}_i^{1/2} \mathbf{w} \right\|$$

$$\|\mathbf{w}\| \leq W$$

$$\xi_i \geq 0$$

where ξ_i are slack parameters of lost. The optional parameter $W \in (0,\infty)$ ensures existence of a solution. When $W = 0$ we lose the control of the parameters of lost. Otherwise, when W is excessively large, $\xi = 0$ and in the case when the data are not linearly separable the solution is not find.

This nonlinear convex optimization problem is solved by interior point method using the Matlab toolbox for optimization over symmetric cones SeDuMi [6]. The decision rule for new subjects characterized by measured values $\boldsymbol{x}$ is $\hat{y}(\boldsymbol{x}) = \operatorname{sign}\left(\langle \mathbf{w}, \boldsymbol{x} \rangle + b\right)$.

Simulation study

In the simulation we classified subjects characterized by different number of selected compounds N, where we started from the most statistically significant compound and step by step we added the next compound in the row. Subjects were 100 times divided into a training set and a testing set (3:2). For each step we estimated the Youden index $J = Se + Sp$ -1 from 100 values of sensitivity and specificity of classification of subjects based on the smoking habit

$$\hat{Se} = \frac{TP}{n^{(1)}} = \frac{\#\{i, y_i = \hat{y}_i | y_i = +1\}}{\#\{i, y_i = +1\}} \qquad \text{and} \qquad \hat{Sp} = \frac{TN}{n^{(2)}} = \frac{\#\{i, y_i = \hat{y}_i | y_i = -1\}}{\#\{i, y_i = -1\}},$$

where TP releases to correct classified subjects from testing set of positive group and TN to correct classified subjects from testing set of negative group. The optional parameters of the robust classification method were chosen from previous studies as $W = 10$ and $\gamma = 0.1$.

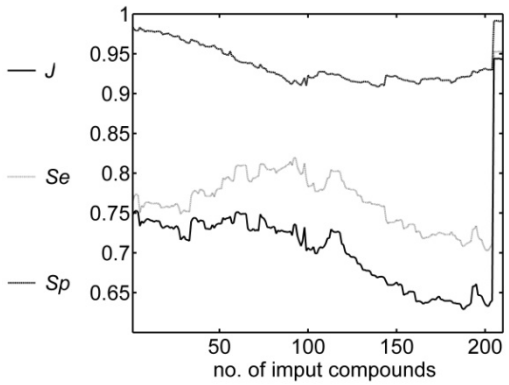

Fig.4. The Youden index *J*, sensitivity *Se* and specificity *Sp* for classification of subjects into the class of smokers or nonsmokers based on different number of breath compounds *N* which were added based on the effectiveness of classification.

Results

The result of the simulation study is visible in Fig. 4. We see that sensitivity and specificity have inverse progress. While specificity is decreasing to N=93, sensitivity (the more important measure for diagnosis) is increasing. Up to N=93 the Youden index remains relatively unchanged. After the point N=93 the observed measures have opposite character, until the point N=205. After this point all observed measures are rapidly increasing.

Discussion

We expected that input data described by more compounds would be better classified to the appropriate classes. In the described classification method we see that the effectiveness of classification is first decreasing. A turning point is appears for higher dimensional data. We think that the reason is that in this higher dimensional space the classes are better linearly separable.

Conclusions

Unlike traditional classification methods the robust classification method assumes that the input data are corrupted with noise. From the result we see, that this classification method gives better result for input data described by more compounds, in other words the robust classification method is optimal for high dimensional data.

Acknowledgements

The research was supported partly by the Scientific Grant Agency of the Slovak Republic (VEGA), grant 2/0019/10 and 1/0077/09 and by the Slovak Research and Development Agency (APVV), grant SK-AT-0003-08.

References

[1] Amann, A., et. al. Model Based Determination of Detection Limits for Proton Transfer Reaction Mass Spectrometer. *Measurement Science Review*, 10 (6), 2010, 180-188.

[2] Bajtarevic, A., et. al. Noninvasive Detection of Lung Cancer by Analysis of Exhaled Breath. *BMC Cancer*, 9 (348), 2009, 1-16.

[3] Bhattacharryya, Ch. Robust Classification of Noisy Data Using Second Order Cone Programming Approach. *Proceeding of Intelligent Sensing and Information Processing*, 2004, 433-438.

[4] Fluss, R., el.al. Estimation of the Youden Index and its Associated Cutoff point. *Biometric Journal*, 47, 2005, 458-472.

[5] Hall, P.G., et. al. Nonparametric Confidence Intervals for Receiver Operating Characteristic Curves. *Biometrika*, 91 (3), 2004, 743-750.

[6] Sturm, J.F. Using SEDUMI 1.02, a Matlab*toolbox for Optimization over Symmetric Cones, *Optimization Methods and Software*, Vol. 11, 1995, 625-653.