MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

A Method for Categorizing Continuous Prognostic Variables Based on the Overall C Index

Harukazu Tsuruta[a], Leon Bax[b]

[a]*Department of Medical Informatics, School of Allied Health Sciences, Kitasato University*
[b] *Department of Medical Informatics, Graduate School of Medical Science, Kitasato University*

Abstract

We propose a new method for categorizing continuous prognostic variables based on the overall discrimination index C introduced by Harrel. We show that this index is closely related to the area under the ROC curve for the original continuous prognostic variable and that the resulting categorized variables have predictive properties comparable to the original variable. However, like some existing approaches, the simple application of our method overestimates the predictive performance. To solve this problem, we propose a parametric method that minimizes this bias and assess its performance with the use of Monte Carlo simulation. The simulation shows that the parametric method is essentially unbiased for both the estimates of predictive performance and the cutoff points, and is a valid tool to create probability profile tables from prediction models for clinical practice.

Keywords:

projections and predictions, prognosis, cutoff point, sensitivity and specificity, categorization

Introduction

Multivariable regression models are useful for both diagnosis and prognosis. If we categorize all continuous independent variables, we can replace the risk function by a table of risks for all possible patient profiles, which is much easier to apply at outpatient clinic [1]. To achieve this, we need a way of optimal categorization. Although many methods have been proposed to dichotomize predictive variables, there has been no integrated method for polychotomization. In this study, we propose a theoretical and practical method for categorization of continuous independent variables.

Methods

For the categorization of continuous variables, we need to (i) define a measure for the predictive ability of prognostic variable with discrete values and then (ii) find the categorization in which the predictive ability is maximized.

We adopted the overall discrimination index C, or the pair consistency probability, introduced by Harrel [2] as a measure of the predictive performance of a categorized independent variable. The C-index is defined as the probability of the correct ranking of a normal-disease-pair. For instance, when we take a sample x_i from the normal group N and another sample y_i from the disease group D randomly; we can call this pair *concordant* if $r(x_i)<r(y_i)$, *tied* if $r(x_i)=r(y_i)$, and *discordant* if $r(x_i)>r(y_i)$, where $r()$ is the rank of the observed value among all values. The ranks are expected to be higher for the diseased cases. The C-index is then defined by

$$C = p_{\text{con}} + p_{\text{tied}}/2 \qquad (1)$$

where

$$p_{\text{con}} = \text{P}(r(x_\text{i})<r(y_\text{i})) \text{ and } p_{\text{tied}} = \text{P}(r(x_\text{i})=r(y_\text{i}))$$

Let α denote the probability that x_i is greater than a cutoff point z for dichotomization, and let β denote the probability that y_i is less than z, i.e.

$$\alpha = \int_z^\infty f_N(x)dx = Fp \qquad \beta = \int_{-\infty}^z f_D(x)dx = 1 - Tp$$

Then the probability of the concordant case becomes

$$p_{\text{con}} = (1-\alpha)(1-\beta),$$

and that of the tied case becomes

$$p_{\text{tied}} = (1-\alpha)\beta + \alpha(1-\beta)$$

Assigning these probabilities into (1), we have

$$C = 1 - (\alpha + \beta)/2 \qquad (2)$$

Since sensitivity is $(1-\beta)$ and specificity is $(1-\alpha)$, the C-index becomes

$$C = (\text{sensitivity} + \text{specificity})/2 \qquad (3)$$

Therefore the highest pair consistency probability is achieved when the sum of two kinds of errors, $\alpha + \beta$, is minimized, or the sum of sensitivity and specificity is maximized.

We can apply the definition of the C-index (1) for polychotomouse cases, and define the *ROC line graph* by joining the points (Fp, Tp) [3]. We can then proof that the C-index is equivalent to the area under the *ROC line graph* and the produced polychotomized variable's predictive

performance is comparable to the original continuous variable.

If we repeat the evaluation of the *C*-index to find optimal cutoff points, for instance by increasing the possible value of the cutoff point with a certain step (*repeated search method*), it overestimates the predictive performance. We therefore developed a practical parametric method in which

1. the distributions of the independent variable in healthy and disease groups are estimated;
2. the optimal cutoff points are found based on the estimated distributions;
3. the predictive ability of the categorized variable is evaluated based on the produced cutoff points.

We assessed its performance with the use of Monte Carlo simulation, both for cases where independent variables are normally distributed and cases where we cannot transform the distributions to normal distributions.

Results and discussion

The simulation shows that the parametric method is essentially unbiased for the estimates of the cutoff points and predictive performance (Figures 1 and 2). We repeated the simulations for various n_h and n_d (n_h=n_d) . Figure 3 shows that the estimation by the parametric method is almost unbiased even if the sample size is relatively small, whereas the naïve repeated search method shows non-negligible bias even when the sample size is large (n =300).

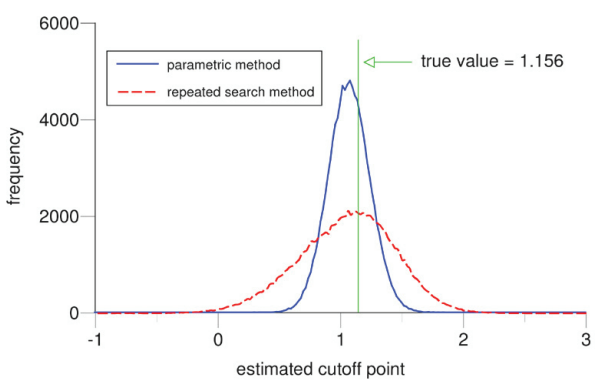

Figure 1 - The frequency distributions of the estimated optimal cutoff point in 100,000 simulations of dichotomization, with $f_h \sim N(0, 1^2)$, $f_d \sim N(2, 1.5^2)$ and n_h=n_d=30

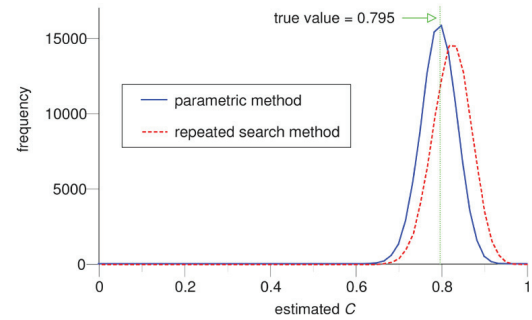

Figure 2 - Distributions of estimated pair consistency probability C in 100,000 simulations of dichotomization

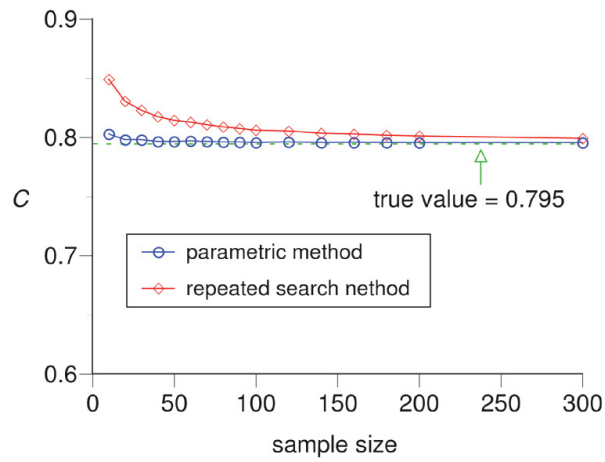

Figure 3 - Changes of estimated pair consistency probability C in dichotomization as a function of sample size

The results for trichotomization and categorization to four classes have a similar tendency for cases where independent variables are normally distributed [3] as well as for cases where we cannot transform the distributions to normal distributions [4].

If a certain type of distribution cannot be assumed, approximation of the distribution curve by a line graph or a restricted cubic spline function creates a workable situation.

Conclusion

We proposed a categorization method for continuous prognostic variables based on the overall discrimination index *C*. We evaluated a parametric method for this problem and showed that it performs well.

References

[1] Tsuruta H, Tsutsumi K, Mochizuki M: Table Presentation of the Risk of Rhabdomyolysis by the Use of an Optimal

Categorization Method for Prognostic Factors and Logistic Regression Analysis. *Proceedings of 11th World Congress on Medical Informatics*, 2004.

[2] Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA: Evaluating the yield of medical tests. *JAMA* 1982, 247(18):2543-2546.

[3] Tsuruta H, Bax L: Polychotomization of continuous variables in regression models based on the overall C index. *BMC Medical Informatics and Decision Making* 2006, 6(41):1-13.

[4] Tsuruta H, Bax L: A criterion and practical method for categorizing continuous predictor variables. *Jap J Med Info*, 2007 (accepted and in print).

Address for correspondence

Harukazu Tsuruta, Ph.D. School of Allied Health Sciences, Kitasato University, 1-15-1 Kitasato, Sagamihara, Kanagawa, Japan, 228-8555. email: ts@med.kitasato-u.ac.jp. Fax: 81-42-778-9799.