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

Harukazu Tsuruta, Leon Bax

Department of Medical Informatics, Kitasato University, Japan

Kitasato University
Medical Informatics
北里大学
医療情報学

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

1. 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 (see the next slide for its definition).
2. We then derived a method for polychotomization mathematically and proved that the *C*-index is closely related to the area under the ROC curve and the produced categorized variable's predictive performance is comparable to the original continuous variable.
3. Since the naïve application of our method, like some existing approaches, overestimates the predictive performance, we developed a parametric method that minimizes this bias and assessed its performance with the use of Monte Carlo simulation.

Pair Consistency Probability *C*

- Measure for the Predictive Ability of a Prognostic Variable -

Let x_i be a sample from the normal group *N*,

and y_i be a sample from the disease group *D*.

We call this pair

concordant	if $r(x_i)<r(y_i)$,
tied	if $r(x_i)=r(y_i)$,
discordant	if $r(x_i)>r(x_i)$,

where r() is the rank of the observed value among all values and is expected to have a larger value for the diseased cases.

Pair Consistency Probability C is then defined by

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

where

$$p_{\mathrm{con}} = P(r(x_i)<r(x_i)) \text{ and } p_{\mathrm{tied}} = P(r(x_i)=r(x_i)).$$

What we have to do next is to find a combination of cutoff points that maximizes the pair consistency probability *C*.

Optimal Cutoff Point for Dichotomization

Let α denote the probability that x_i is greater than a cutoff point z, and let β denote the probability that y_i is less than z, then the probability of the concordant case becomes

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

and that of the tied case becomes

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

Consequently,

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

Since *sensitivity* is $(1 - \beta)$ and *specificity* is $(1 - \alpha)$,

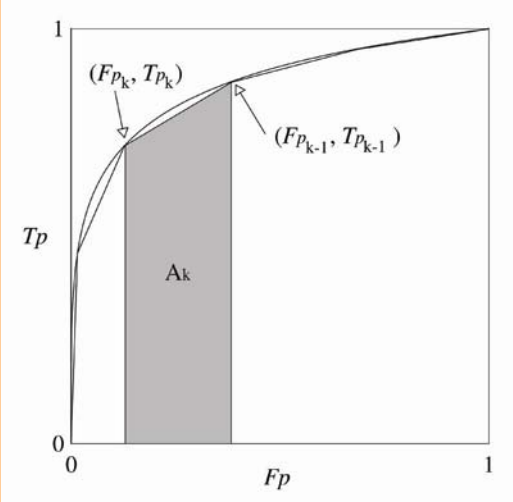
$$C = 1 - (\alpha+\beta)/2 = (\textit{sensitivity} + \textit{specificity}) / 2. \qquad (2)$$

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

Extension to Polychotomization

We can also apply the definition of the *C*-index (1) for polychotomouse cases, and define "optimal" cutoff points by maximizing the *C*-index.

We then define the *ROC line graph* by joining the points (*Fp*, *Tp*), each corresponding to one of the cutoff points for polychotomization. We can then proof that the *C*-index is equivalent to the area under the *ROC line graph* and that the produced polychotomized variable's predictive performance is comparable to the original continuous variable [3].



NB: *Fp* = false positive fraction, *Tp* = true positive fraction

Kitasato University
Medical Informatics
北里大学
医療情報学

Parametric Method for Categorization

If we repeat the evaluation of *C* 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. To solve this problem, we developed a practical parametric method in which

(i) the distributions of the independent variable in healthy and disease groups are estimated;

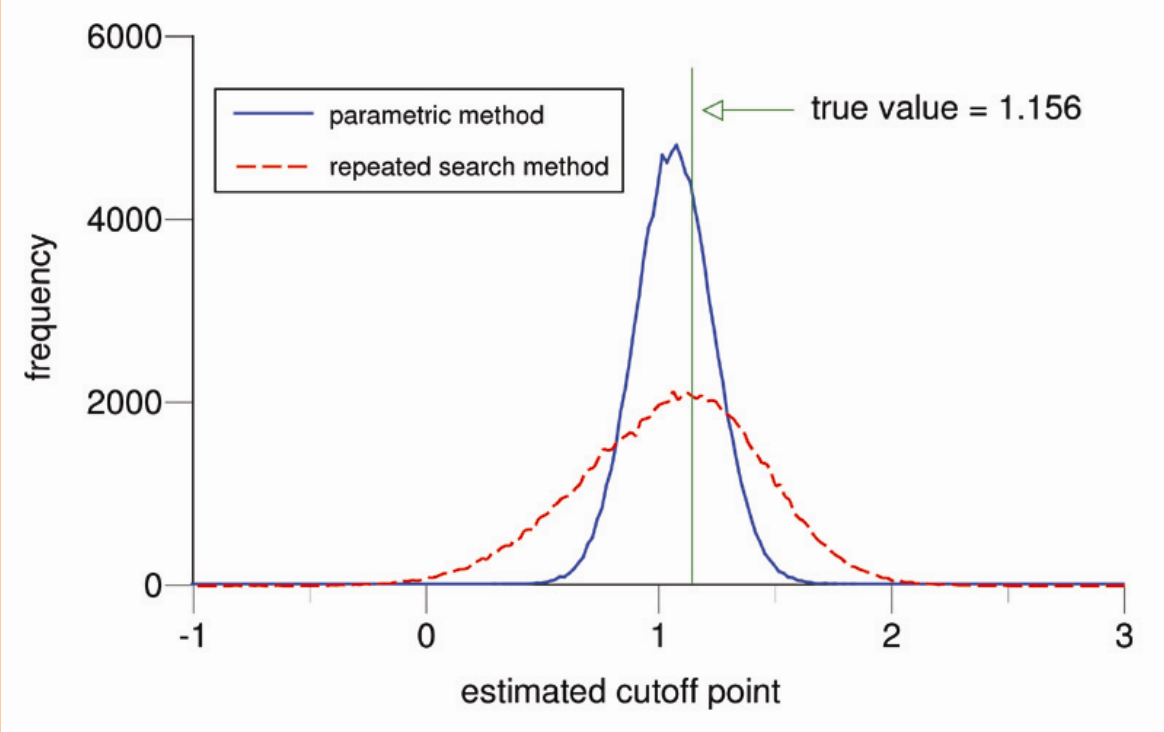
(ii) the optimal cutoff points are found based on the estimated distributions;

(iii) 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 the independent variables are normally distributed and for cases where we cannot transform the distributions to normal distributions.

Results

The frequency distributions of the estimated optimal cutoff point in 100,000 simulations of dichotomization

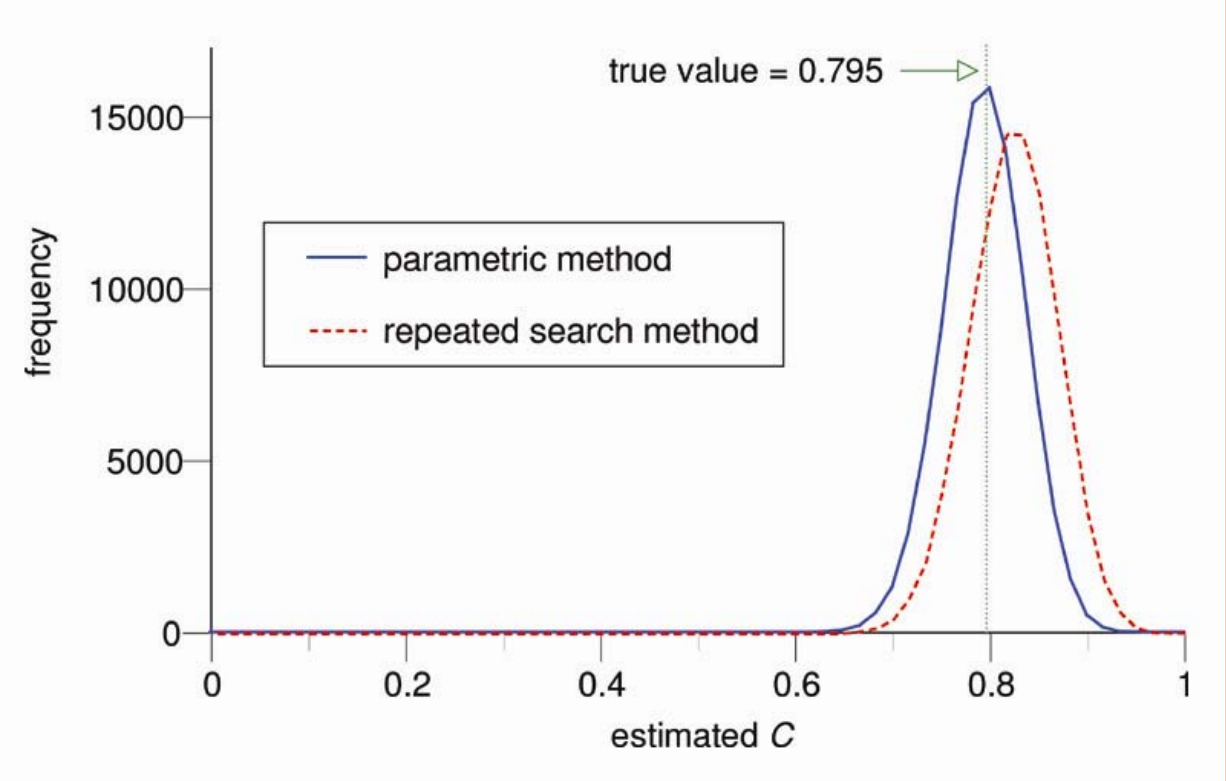


The parametric method provides a estimator for the cutoff point with higher precision.

($f_h \sim N(0, 1^2)$, $f_d \sim N(2, 1.5^2)$ and $n_h = n_d = 30$)

Kitasato University
Medical Informatics
北里大学
医療情報学

Distributions of estimated pair consistency probability *C* in 100,000 simulations of dichotomization

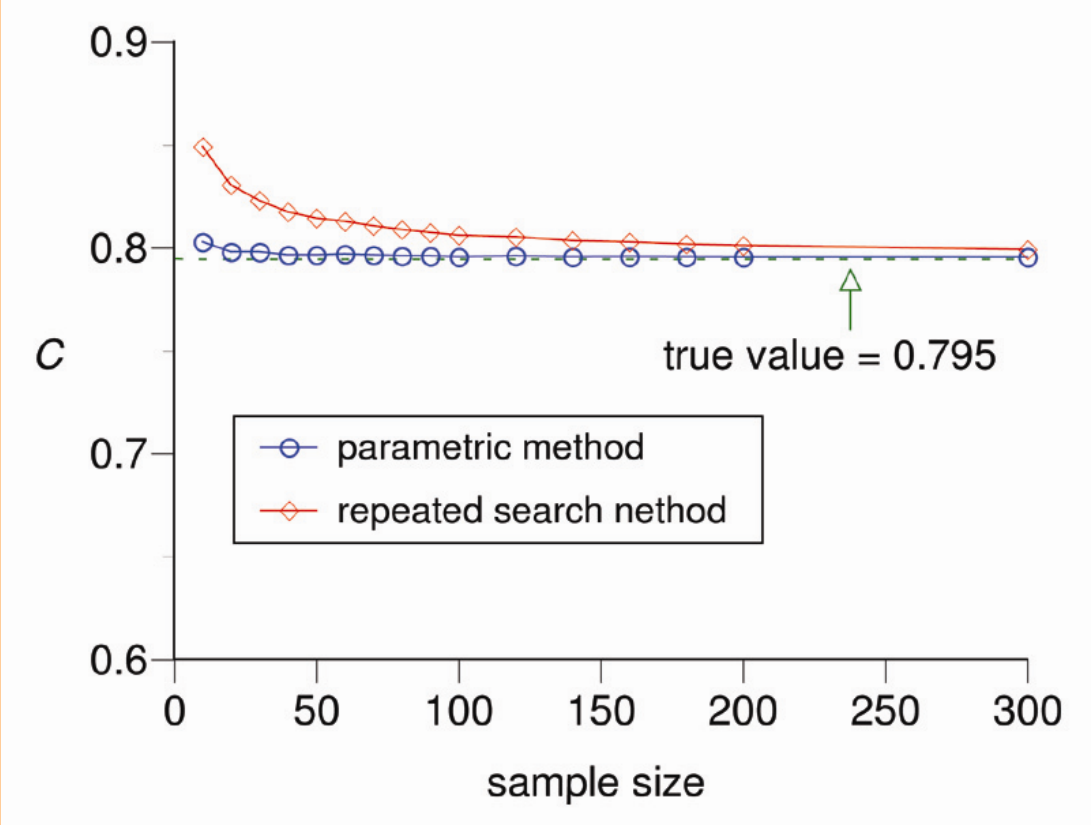


The parametric method is essentially unbiased for the predictive performance.

($f_h \sim N(0, 1^2)$, $f_d \sim N(2, 1.5^2)$ and $n_h = n_d = 30$)

Kitasato University
Medical Informatics
北里大学
医療情報学

Changes of estimated pair consistency probability *C* in dichotomization as a function of sample size



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).

Kitasato University
Medical Informatics
北里大学
医療情報学

Discussion

The results for trichotomization and categorization to four classes have a similar tendency, both for cases where independent variables are normally distributed [3] and for cases where we cannot transform the distributions to normal distributions [4].
If we cannot assume a certain type of distribution, 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.

Kitasato University
Medical Informatics
北里大学
医療情報学

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.