Discrimination and calibration are concurrently required for model comparison

Jainn-Shiun Chiu, Fu-Chiu Yu, Yu-Chuan Li *

Graduate Institute of Medical Informatics, Taipei Medical University, No. 250, Wusing St., Sinyi District, Taipei City 110, Taiwan

Received 26 May 2005; accepted 24 July 2005
Available online 26 October 2005

Keywords: Bayesian classifier; Artificial neural network; Machine learning

To the Editor:

We read with interest the research article by Bigi et al. [1], and praise their achievement to build two machine learning models, artificial neural network and Bayesian classifier, and compare their performance for classifying predischarge risks in patients with uncomplicated myocardial infarction on the basis of exercise electrocardiography and pharmacological stress echocardiography.

In the field of machine learning, artificial neural network model is an artificial intelligence composed of individual nonlinear processing elements arranged in highly interconnected layers based on the paradigm of biological nervous system. It has been increasingly applied as a revolutionary model in clinical medicine with the help of advances in computer-aided analysis. On the other hand, Bayesian classifier model is based on the Bayesian theorem principally developed for performing classification tasks. Bayesian classifier model supposes that the input variables are statistically independent. It is a particularly appropriate easy-to-use classification tool when the number of dimensions of the input variables is high and can often outperform some complicated methods. Therefore, the quality of the comparative analysis between artificial neural network and Bayesian classifier models should be paid more attention. To increase the quality for classification model in clinical research, it would be more proper to calculate *discrimination* and *calibration* concurrently [2]. Discrimination is a measure of how well a model to recognize subjects correctly as two different classes. From the perspective of goodness-of-fit, calibration evaluates the degree of correspondence between the estimated probabilities produced by a model and the actual observation.

Common assessments used in discrimination for predictive classification include sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios for positive and negative tests, and area under the receiver–operating characteristics curve [3]. In contrast, although many investigators used area under the receiver–operating characteristics curve with the best simultaneous sensitivity and specificity to determine discriminatory power of a model, a good discrimination has the probability of poor calibration when classification outputs are metamorphosed monotonically [4]. To avoid this pitfall, calibration using misclassification rate, Pearson's χ^2, or Hosmer–Lemeshow statistic [5] should be considered. In addition, inter-rater agreement with kappa value could be adopted to approach the reproducibility and repeatability [6].

The authors concluded that their artificial neural network model did not perform better than Bayesian classifier model by demonstrating the sensitivity and specificity without mentioning area under the receiver–operating characteristics curve, which can provide a superior index of the discrimination for each model. They also overlooked other useful calibration assessments for comparing the goodness-of-fit statistic of their models. According to their results, readers could not identify which model is truly better. In the era of evidence-based medicine, new predictive model should be carefully and critically appraised since inadequate evaluations may lead to wrong conclusions.

* Corresponding author. Tel.: +886 2 23776730; fax: +886 2 27339049.
E-mail address: jack@tmu.edu.tw (Y.-C. Li).

0167-5273/$ - see front matter
doi:10.1016/j.ijcard.2005.07.038

References

[1] Bigi R, Gregori D, Cortigiani L, Desideri A, Chiarotto FA, Toffolo GM. Artificial neural networks and robust Bayesian classifiers for risk stratification following uncomplicated myocardial infarction. Int J Cardiol 2005;101:481–7.

[2] Li YC, Liu L, Chiu WT, Jian WS. Neural network modeling for surgical decisions on traumatic brain injury patients. Int J Med Inform 2000; 57:1–9.

[3] McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Mak 1984;4:137–50.

[4] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352–9.

[5] Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92–106.

[6] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.