Discussion

Class probability estimation for medical studies

Richard Simon*

Biometric Research Branch, National Cancer Institute, 9609 Medical Park Drive, Rockville, MD20892-9735, USA

Received 18 December 2013; revised 18 December 2013; accepted 19 December 2013

I provide a commentary on two papers “Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory” by Jochen Kruppa, Yufeng Liu, Gérard Biau, Michael Kohler, Inke R. König, James D. Malley, and Andreas Ziegler; and “Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications” by Jochen Kruppa, Yufeng Liu, Hans-Christian Diener, Theresa Holste, Christian Weimar, Inke R. König, and Andreas Ziegler. Those papers provide an up-to-date review of some popular machine learning methods for class probability estimation and compare those methods to logistic regression modeling in real and simulated datasets.

Keywords: Logistic regression; Machine learning; Prediction.

1 Introduction

In the past 15 years, breakthroughs in biotechnology, particularly gene expression arrays, have stimulated statisticians and computer scientists to address $p > n$ classification problems where the number of candidate features (variables) p exceeds the number of subjects n. This development has been particularly important for statistics, which has overly focused on inference and statistical significance testing, leading many scientists to seek out computer scientists, physicists or engineers who sometimes seem more open to other approaches. Large literatures have developed in both the statistical and machine learning journals on new methodology for high dimensional classification. Many medical problems are not really classification problems, however, they are decision problems. For example, the disease may either be confined or disseminated; if it is confined then the preferred therapy is X, whereas if it is disseminated the preferred therapy is Y. A given medical test (T) may provide information about whether the disease is confined or disseminated, but that information is often imperfect. So, rather than classifying the patient as having confined or disseminated disease based on T, it might be better to provide a numerical measure (probability) for the disease status. That way, the patient and physician can make a decision about treatment selection that is appropriate for the utilities of that patient. The problem of estimating class probabilities addressed by the two papers by Kruppa et al. (2014a,b) is very important in medicine.

The distinction made between statistical methods and machine learning methods is really a distinction between methods developed for inference and methods developed for discovery, classification, and prediction. In selecting a method for an application it is important to be clear about the objective of the application and methods should be evaluated based on their intended use. For example, early phase clinical trials may be attempting to discover which baseline biomarkers are predictive of response to a treatment. Phase III clinical trials for regulatory approval focus more on inference about a treatment effect in a defined population. Inferential methods are generally ill suited for problems of discovery,

*Corresponding author: e-mail: rsimon@mail.nih.gov, Phone: +1-240-2766028 prediction or classification and machine learning methods are generally ill suited for problems of inference. Statistics as a field should, however, encompass more than inferential methods, certainly in the era of “data science”.

2 How should class probability estimators be evaluated?

The theoretical paper (Kruppa et al., 2014a) focuses heavily on consistency of the class probability estimator. The simulations in the theoretical paper clearly establish, however, that consistency is not a useful guide for selecting a class probability estimator. The simulations show many examples of very poor accuracy for consistent probability estimators even in cases with very small p and very large n. Certainly with moderate to large p, medical studies rarely have an n that puts us anywhere near asymptopia for estimating a function of unknown form.

The simulation studies compared the class probability estimates to the true class probabilities for a sample of feature vectors. This showed that for these $p \ll n$ problems, accurate reconstruction of the unknown true class probability function was rarely achieved.

For real data, one cannot compare the estimated class probability function with the true function. One can compute the Brier score (the average squared error between the class indicator and the estimated class probability) in the validation set. If one does not have a separate validation dataset, one can compute a cross-validated Brier score (Kim and Simon, 2010) in the full dataset. The applications paper (Kruppa et al., 2014b) also uses the area under the receiver operating characteristic curve (AUC) to evaluate the estimated class probability functions for the real data examples. Although the AUC does measure the correlation between the probability estimate and the class indicator in two-class problems, it is more appropriate for comparing classifiers, not for comparing class probability estimators. The limitations of the AUC have been discussed by the authors (Kruppa et al., 2012) and by others (Pepe and Janes, 2008).

Accurately estimating the true class probability function is very difficult except in simple circumstances with very small p, very large n and restriction of functional forms. This was shown previously (Kim and Simon, 2010) in a study of high dimensional class probability estimation. Accurate estimation of the class probability function is more difficult than accurate classification. A more basic requirement than accurate reconstruction of the true class probability function, however, is that the estimated class probability function should provide well-calibrated forecasts (3). This means that of future cases estimated to have a probability of being in class A of about p, about $100p\%$ will actually be in class A and this should be true for all p in (0,1). For a two class problem let $\hat{f}(x)$ denote the estimated probability that a case with feature vector x is in class A and let $f(x)$ denote the true class probability function. For any $p \in (0, 1)$ let $X(p) = \{x \ni \hat{f}(x) = p\}$. Then a forecast is well calibrated if for all $p \in (0, 1)$, the average of $f(x)$ over the set $X(p)$ is p. For a real application, if one has a separate validation set, one can compute $\hat{f}(x)$ for each case in the validation set, and plot p versus a window estimator of the proportion of cases with $\hat{f}(x) \in (p - w, p + w)$ that are in class A. In the validation set one can also fit a logistic regression of the indicator of whether a case with feature vector x is class A versus $\hat{f}(x)$ and test whether the slope is 1 and the intercept 0. A logistic regression classifier can potentially be properly calibrated in this way. Calibration can also be evaluated using cross-validation if there is not a separate validation set (Kim and Simon, 2010).

If one is proposing use of a quantitative score function as a class probability estimator, then one should demonstrate that the function estimated based on the finite sample of cases at hand is well calibrated for future use. It is of no real relevance whether the class of quantitative score functions is asymptotically well calibrated, the specific estimator function developed on the finite training set must be well calibrated for new cases. Although there is no automatic way of adjusting most machine learning class probability estimates to make them well calibrated, it is important that they be approximately well calibrated before they are recommended for use.

Providing well-calibrated forecasts is a minimal requirement for a medically useful class probability estimator. Kim and Simon (2010) also discuss refinement for selecting among well-calibrated function estimators. Refined class probability estimators have large variance of $\hat{f}(x)$ with regard to the distribution of the x vector. Several methods of class probability estimation such as L1 and L2 penalized logistic regression, and naïve Bayes shrunken centroid estimators were evaluated by Kim and Simon (2010) with regard to calibration, refinement, and approximation to the true class probability function for $p > n$ settings.

3 Acceptability of the class probability estimator in medical practice

There is an enormous gap between the large literature on prognostic models and the small number of models used in medical practice. There are several reasons for this discrepancy. Most prognostic models do not provide actionable information. That is, they are based on analysis of a heterogeneous set of patients who received a variety of treatments. Physicians want tools that help them make treatment decisions. Unless that decision context is clearly and specifically defined at the outset of the study and used to drive the selection of the training set, the resulting model is unlikely to find medical acceptance.

There are several other reasons for lack of model acceptance but one is that addressed in the discussion section of the applications paper (Kruppa et al., 2014b). Interpretability and ease of use are important for acceptance of a multivariate test result. For medical studies, there is resistance of pathologists and medical practitioners to use uninterpretable black box multivariate tests for making important treatment decisions. Parsimonious models that can be presented as nomograms are increasingly popular in the medical literature (Kattan, 1998). If those features make biological and medical sense and the test provides well-calibrated forecasts that help patients and physicians make important medical decisions, then acceptance is more likely.

The nearest neighbor and random forest class probability estimators lack interpretability and convenience unless the number of features and number of trees are very limited. A great attraction of CART (Breiman et al., 1984) was the interpretability of a single decision tree. Forests avoid the instability of single trees, but at the cost of interpretability. This can be remedied by identification of representative trees (Banerjee et al., 2012). The logistic regression models developed in the applications paper were convenient and interpretable and were either best or competitive with the best in terms of the evaluation measures used in all of the applications studied except for the thyroid example. It is possible, however, that better feature selection would have made logistic regression competitive with the random forest on the thyroid dataset. In the applications, feature selection was based on stepwise regression using very stringent nominal statistical significance thresholds. Feature selection should, however, be optimized using penalized regression methods for minimizing the cross-validated Brier score. Conventional logistic feature selection is based on a less relevant resubstitution measure of nominal statistical significance. Quadratic terms and two-way interactions can also be included in penalized logistic regressions as candidate variables using the group lasso (Meier et al., 2008).

4 Final points

The performance of the random forest class probability estimator for the thyroid disease example appeared to provide exceptional classification and it warrants further study to see whether as a class probability estimator it is well calibrated. It would also be of interest to better understand the function it is computing and whether that function could be approximated more simply and parsimoniously. It is surprising that a random forest of 10 000 trees constructed from mtry = 3 randomly selected features would provide such good performance. That approach would not be expected to work well in $p > n$ settings where most of the features are noninformative.

The poor performance of logistic regression on a couple of the highly nonlinear simulated datasets in (Kruppa et al., 2014a) may be an anomaly that should not be too great a cause of concern. With such large *n*, quadratic or higher terms could easily be included in the model or included as candidate variables in a penalized logistic regression.

I congratulate the authors on providing two excellent papers that contribute in several important ways to development and use of class probability estimators. These papers provide an excellent introduction to some methods that many readers may not be familiar with, provide an extensive review of research on those methods, and provide a wealth of extensively analyzed examples for evaluation of those methods on simulated and real datasets. Future extensions might include other potentially valuable machine learning methods and feature selection techniques.

The authors should also be congratulated on presenting an evaluation of methods free from the pitfalls that so often are encountered in publications evaluating classification models. These pitfalls involve implicitly using the validation set for feature selection (Simon et al., 2003) or for tuning parameter optimization (Varma and Simon, 2006). The authors apparently optimized their tuning parameters on the training set without using resampling and this may lead to excessively over-fit models. I typically use 10-fold or fivefold cross validation on the training set to optimize tuning parameters. I then use the optimally selected tuning parameter in fitting the model to the full training set and use that model to estimate class probabilities for the validation set. This is similar to the double bootstrap method recently proposed by the authors (Kruppa et al., 2013). Since we are rarely close to asymptopia in practice, some degree of over-fitting is necessary for accurate prediction, but excessive over-fitting can badly degrade performance (Kim and Simon, 2014).

Conflict of interest

The author has declared no conflict of interest.

References

Banerjee, M., Ding, Y. and Noone, A. M. (2012). Identifying representative trees from ensembles. *Statistics in Medicine* **31**, 1601–1616.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regressin Trees*. Wadsworth: Belmont, CA.

Kattan, M. W., Eastham, J. A., Stapleton, A. M. F., Wheeler, T. M. and Scardino, P. T. (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *Journal of the National Cancer Institute* **90**, 766–771.

Kim, K. and Simon, R. (2010). Probabilistic classifiers with high dimensional data. *Biostatistics* **12**, 399–412.

Kim, K. and Simon, R. (2014). Overfitting, generalization and MSE in class probability estimation with high-dimensional data. *Biometrical Journal* **56**, 256–269.

Kruppa, J., Ziegler, A. and Konig, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. *Human Genetics* **131**, 1639–1654.

Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A. (2013). Consumer risk: Individual probability estimates using machine learning. *Expert Systems with Applications* **40**, 5125–5131.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D. and Ziegler, A. (2014a). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal* **56**, 534–563.

Kruppa, J., Liu, Y., Diener, H. C., Holste, T., Weimar, C., König, I. R. and Ziegler, A. (2014b). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal* **56**, 564–583.

Meier, L., De Geer, S. and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B* **70**, 53–71.

Pepe, M. S. and Janes, H. E. (2008). Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *Journal of the National Cancer Institute* **100**, 978–979.

Simon, R., Radmacher, M. D., Dobbin, K. and McShane, L. M. (2003). Pitfalls in the analysis of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **95**, 14–18.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91.