JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

RICHARD D. DE VEAUX

Discussion and comments. Data mining et statistique

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 53-58

<http://www.numdam.org/item?id=JSFS_2001__142_1_53_0>

DISCUSSION AND COMMENTS

Data Mining et Statistique

Richard D. DE VEAUX *

1. Introduction

I would like to start by thanking the authors for their overview of data mining. As they point out so clearly, there are many areas of intersection between data mining and statistics. In fact, what exactly is the difference between the two? I am reminded of an exchange that occurred at a recent conference. After his talk on boosting, Jerry Friedman was asked whether there is any difference between data mining and statistics. He asked, "Do you want the short answer?". After receiving an affirmative response, he said, "OK, no". Of course, the longer answer to the question is more complex. There are differences of intention, of points of view and of the backgrounds of the users of data mining and statistics. I would like to focus on these different points of view and in particular on the willingness of data miners to combine models to achieve better prediction.

2. Interpretability

Besse *et al* point to automation of the modeling process as one of the differences between statistics and data mining. Statisticians, generally, spend quite a bit of effort on their choice of model. They care about their models. Some might argue even too much. George Box has joked that "statisticians, like artists, have the bad habit of falling in love with their models". They carefully select the predictor variables they will enter, or possibly enter into the model, and think hard about the error structure of their models. They also want the model to be interpretable and to make sense. For them, an ideal model contains at most a few variables, perhaps an 2-way interaction of one of the pairs, and is additive in these terms, if not linear. By contrast, data miners tend to automate the process, viewing as unnecessary and time

* Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267 U.S.A. Visiting Professor, Laboratoire Probabilité et Statistique, Université Paul Sabatier, Toulouse, 31000, FRANCE, e-mail rdeveaux@williams edu

consuming much of the fussing about the model. The main goal for them is a model that predicts well.

Of course, there are times that interpretability is essential. A bank may not be able to use a black box model that simply predicts whether someone is a good credit risk or not without some explanation of why. An applicant may have a right to know more about why their application was rejected than, "the neural network told us you were a bad risk". Many of the data mining models, notably neural networks and support vector machines are "black box models". These methods are powerful function fitters, but provide no easy way to see what is going on inside the box. For neural networks, a large literature has been devoted to attempts to find interpretable networks given a large, overparameterized network. An attempt to do model selection by eliminating some of the connections (setting weights to zero) is known by the term Optimal brain damage (LeCun, Denker and Solla [1990]), or optimal cell damage (Cibas *et al.* 1994) due to the analogy with the brain. Generally, these methods have not been terribly successful due to the great number of possible smaller networks contained within a larger one. Recently however, some progress has been made in finding understandable explanations for a general "black box" model. See Owen (2000,1994,1992), Friedman (2001), An and Owen (2001) and LeMieux and Owen (2001). This work may help to bridge the gap between the two kinds of models.

3. Combining models

Forecasters have long known that combining predictions by averaging predictions from different forecasting models leads to improvement in the forecasts (see Bates and Granger [1969], Newbold and Granger [1974] for example). Averaging can help both by reducing the biases produced by different classes of models and by decreasing the variance of the predictions. Recently, two new methods for combining models have appeared in the literature and attracted much attention in the data mining community. These methods, known as bagging and boosting, are both based on model averaging, but they do it in very different ways. Both approaches take a simple model, called a weak learner, as their base. Then, many different versions of this model are created. The ensemble of generated models is called a committee. To predict the value of the response for each data point, an average, or majority vote, or some other linear combination of the committee is taken.

Often the weak learners on which the methods are based are decision trees. These models are popular both among statisticians and the machine learning community, because they are simple, often very interpretable, and probably also because they were discovered at about the same time by both communities. We'll illustrate how both bagging and boosting work by using trees as the weak learner.

Bagging (or bootstrap aggregation) is based on a very simple idea. Take an ensemble of different bootstrapped samples of the data and fit a tree to each

sample. Then, for each data point, simply take a majority vote of the trees to determine the prediction of its class (for classification) or take the average of the predictions for regression. Bagging, like averaging in general, tends to reduce the *variance* of the predictions. Since trees tend to have high variance, this simple idea works quite well. The bias is not necessarily reduced because all the models are similar, in this case trees, which tend to have large bias. Reduced prediction variance is gained, but what is lost? Unfortunately, the resulting model is no longer directly interpretable. For each data value, we can report only the average class or value of the ensemble of trees. The individual trees may even, and quite likely, contain different predictors. Thus the prediction of one tree may be based on a completely different set of predictors than another tree in the same bootstrapped ensemble. It is difficult to judge even the relative importance of the predictor variables in this case.

Boosting, also takes a majority vote of models, but does so in a very different way. Boosting works by sequentially applying the same algorithm to *reweighted* (not resampled) versions of the training data. Like bagging, it then takes a weighted majority vote of the sequence of classifiers (or predictors) thus produced. The basic idea was found in a series of papers by Shapire (1990), Freund (1995) and Freund and Shapire (1997), but for statisticians, a wonderful explanation of why and how boosting works can be found in the paper by Friedman, Hastie and Tibshirani (2000). The basic idea for classification, as shown in the first two papers, is to start with a "weak learner", for example a tree with very few nodes. After fitting the data to this weak learner, one then reweights the data, giving greater weight to points that were misclassified. One repeats this many times, ending up with a series of weak learners that seem to concentrate more and more on points that have previously been misclassified. Originally, like bagging, a majority vote of the weak learners in the "committee" was taken, but in a version called AdaBoost (Freund and Shapire, 1997), a linear combination of the predictions of each weak learner in the "committee" was used in order to produce the final prediction.

What Friedman, Hastie and Tibshirani managed to show is that the AdaBoost algorithm is really a combination of two things. At the core is an additive logistic regression which is repeated in a stage-wise manner. To think about what this means, imagine linear regression for a minute. A step wise procedure would take the single best predictor as the first variable to enter the model. After it is entered, it then takes the next best predictor (conditional on the first being in the model) and *refits both coefficients.* A stage-wise linear regression would start the same way, taking the single best predictor, but, by contrast, would *not change* the coefficient of this variable, but simply find the coefficient for the second variable that was optimal given the first and its particular coefficient were already in the model. So, stage wise refers to the fact that at each step, the original model is not refit, but each prediction is updated by a new model from that point. Friedman (2001) then goes on to generalize this idea, building a general framework that includes AdaBoost as a special case, but which includes a variety of different weak learners and associated loss functions. He finds that in practice, small trees work better than additive

logistic regression. The resulting stage-wise addition of many, many small trees is called MART (multiple additive regression trees). (A version of MART for S-plus is available from http ://www-stat.stanford.edu/ jhf/. For details see Friedman, 1999).

The result from the addition of usually at least several hundred, and often several thousand small trees is that the resulting model is far from directly interpretable. But Friedman describes "partial dependence" functions of the form $B_L(x_L) = (1/n)\sum_{i=1}^{n} B(x_L, x_{i,-L})$ where $L \subseteq \{1, \ldots, d\}$ describes some variables of interest x_L denotes those x-variables in L and $x_{i,-L}$ denotes x_{ij} for $j \notin L$. These variable importance functions help to make the "black box" method interpretable.

There are other methods available for combining models. Jordan and colleagues (Jordan and Jacobs, 1994) have developed a mixture of experts model that uses adaptive methods to combine models internally at different data points. In a very different tack, De Veaux *et al* (1999) showed that in a comparison of data mining methods, a simple combination of applying a linear model first and then a non-linear model to the residuals nearly always outperformed both the linear and non-linear methods, when judged by performance on an independent test set of data.

4. Challenges and Directions

Combining models is an exciting direction for improving prediction. Statisticians should become more involved in both the practice and the research of these methods. Perhaps the recent progress in interpreting these models will help statisticians over their hesitancy. But, there are still challenges looming over data mining. The first is missing data. In a data set with millions of rows (cases) and hundreds of columns (predictors), there are sometimes *no* complete cases. So, the simple practice of row deletion yields a completely empty data set. Knowing how to proceed can be very difficult. Progress in imputation, especially in multiple imputation (Rubin, Schafer), is promising, but, in practice, the problem of missing data for large data sets is still a huge challenge.

Another challenge is the data base itself. With all the model selection available and all the automation to help, there is no guarantee that the right variables have been measured, or if they have that they have been measured accurately. This simple fact is often ignored in the middle of a large data mining project. A related caveat is that an important variable in the model $B(x)$ is not necessarily an important variable in real world prediction. Of course, this problem even arises in linear regression. One might find a good linear model fits using either x_{i1} or x_{i2} without the other. And, no amount of observational data and modeling expertise can establish a causal relationship, or even the direction of a causal relationship, between predictors and response. The oft-repeated slogan in statistics courses is that "correlation does not establish

causation". This point can be forgotten when very sophisticated machine learning models are used.

Data mining is an enormous opportunity for statisticians. There are challenging problems in both applications and research. Basic statistical principles are often ignored. This can be viewed either with horror, or as an opportunity by statisticians. There is much work to be done here. If it is not done by statisticians, it will be done by others. There is a long list of other fields that had their origins in statistics, but were later ignored by us (see Friedman, 1998). Let's not let data mining become the next one on that list.

RÉFÉRENCES

[1] AN J. and OWEN A. B. (2001) Quasi-Regression. Journal of Complexity. To appear.

[2] CIBAS T., SOULIÉ F., GALLINARI P. and RAUDYS S. (1994). "Variable selection with optimal cell damage". In Maria Marinaro and Pietro G. Morasso, editors, *Proceedings of the International Conference on Artificial Neural Networks*, volume 1, pages 727–730. Springer-Verlag.

[3] DE VEAUX R.D., UNGAR L.H. and BAIN R.(1999) "Neural Networks for Chemometrics : A critical evaluation", Talk given at the Fall Technical Conference, American Society for Quality, Houston, October 1999.

[4] FRIEDMAN J. H. (1998) "Data mining and Statistics : What's the difference ?" Technical Report, *Stanford University* available at http ://www-stat.stanford.edu/ jhf/.

[5] FRIEDMAN J. H. (1999) "Tutorial : Getting started with MART in S-plus" Technical Report,*Stanford University* , available at http ://www-stat.stanford.edu/ jhf/.

[6] FRIEDMAN J. H. (2001) "Greedy function approximation : A gradient boosting machine". *The Annals of Statistics* Vol. 29, No. 4.

[7] FRIEDMAN J., HASTIE T. and TIBSHIRANI R. (2000) "Additive logistic regression : a statistical view of boosting (with discussion)", *Annals of Statistics*, 28 : 337-387.

[8] FREUND Y. (1995), 'Boosting a weak learning algorithm by majority', *Information and Computation* 121(2), 256-285.

[9] FREUND Y. and SCHAPIRE R. E. (1997), 'A decision-theoretic generalization of online learning and an application to boosting', *Journal of Computer and System Sciences* 55, 119-139.

[10] JORDAN M. I., and JACOBS R. A.(1994) "Hierarchical Mixtures of experts and the EM algorithm" *Neural Computation* **6**, 181-214.

[11] LECUN Y., DENKER J.S. and SOLLA S.A. (1990) Optimal Brain Damage in *Advances in Neural Information Processsing Systems 2* (D.S. Touetzky, Edl.), San Mateo, CA, Morgan Kaufmann, pp. 598-605.

[12] LEMIEUX C. and OWEN A. B. (2001) "Quasi-Regression and the Relative Importance of the ANOVA Components of a Function" To appear in proceedings of MCQMC-2000, Harald Niederreiter and Fred Hickernell, editors, Springer.

[13] OWEN A. B. (1992) "Orthogonal Arrays for Computer Experiments, Integration and Visualization". Statistica Sinica. Volume 2, Pages 439–452.

[14] OWEN A. B. (1994) "Lattice Sampling Revisited : Monte Carlo Variance of Means Over Randomized Orthogonal Arrays". *Annals of Statistics*, 22, 930–945.

[15] OWEN A. B. (2000) Assessing linearity in high dimensions. *Annals of Statistics*, volume 28, number 1, 1-19.

[16] SCHAPIRE R. E. (1990), 'The strength of weak learnability', *Machine Learning* 5(2), 197-227.