Evaluation of Data Mining Methods

Paolo Giudici
University of Pavia, Italy

INTRODUCTION

Several classes of computational and statistical methods for data mining are available. Each class can be parameterised so that models within the class differ in terms of such parameters (see, for instance, Giudici, 2003; Hastie et al., 2001; Han & Kamber, 2000; Hand et al., 2001; Witten & Frank, 1999): for example, the class of linear regression models, which differ in the number of explanatory variables; the class of Bayesian networks, which differ in the number of conditional dependencies (links in the graph); the class of tree models, which differ in the number of leaves; and the class multilayer perceptrons, which differ in terms of the number of hidden strata and nodes. Once a class of models has been established the problem is to choose the "best" model from it.

BACKGROUND

A rigorous method to compare models is statistical hypothesis testing. With this in mind one can adopt a sequential procedure that allows a model to be chosen through a sequence of pairwise test comparisons. However, we point out that these procedures are generally not applicable in particular to computational data mining models, which do not necessarily have an underlying probabilistic model and, therefore, do not allow the application of statistical hypotheses testing theory. Furthermore, it often happens that for a data problem it is possible to use more than one type of model class, with different underlying probabilistic assumptions. For example, for a problem of predictive classification it is possible to use both logistic regression and tree models as well as neural networks.

We also point out that model specification and, therefore, model choice is determined by the type of variables used. These variables can be the result of transformations or of the elimination of observations, following an exploratory analysis. We then need to compare models based on different sets of variables present at the start. For example, how do we compare a linear model with the original explanatory variables with one with a set of transformed explanatory variables?

The previous considerations suggest the need for a systematic study of the methods for comparison and evaluation of data mining models.

MAIN THRUST

Comparison criteria for data mining models can be classified schematically into: criteria based on statistical tests, based on scoring functions, computational criteria, Bayesian criteria and business criteria.

Criteria Based on Statistical Tests

The first are based on the theory of statistical hypothesis testing and, therefore, there is a lot of detailed literature related to this topic. See, for example, a text about statistical inference, such as Mood, Graybill, & Boes (1991) and Bickel & Doksum (1977). A statistical model can be specified by a discrete probability function or by a probability density function, f(x). Such model is usually left unspecified, up to unknown quantities that have to be estimated on the basis of the data at hand. Typically, the observed sample it is not sufficient to reconstruct each detail of f(x), but can indeed be used to approximate f(x) with a certain accuracy. Often a density function is parametric so that it is defined by a vector of parameters $\Theta=(\theta_1,\ldots,\theta_l)$, such that each value θ of Θ corresponds to a particular density function, $p_\theta(x)$. In order to measure the accuracy of a parametric model, one can resort to the notion of distance between a model f, which underlies the data, and an approximating model g (see, for instance, Zucchini, 2000). Notable examples of distance functions are, for categorical variables: the entropic distance, which describes the proportional reduction of the heterogeneity of the dependent variable; the chi-squared distance, based on the distance from the case of independence; and the 0-1 distance, which leads to misclassification rates. For quantitative variables, the typical choice is the Euclidean distance, representing the distance between two vectors in a Cartesian space. Another possible choice is the uniform distance, applied when nonparametric models are being used.

Any of the previous distances can be employed to define the notion of discrepancy of an statistical model.

The discrepancy of a model, g, can be obtained comparing the unknown probabilistic model, f, and the best parametric statistical model. Since f is unknown, closeness can be measured with respect to a sample estimate of the unknown density f. A common choice of discrepancy function is the Kullback-Leibler divergence, which can be applied to any type of observations. In such context, the best model can be interpreted as that with a minimal loss of information from the true unknown distribution.

It can be shown that the statistical tests used for model comparison are generally based on estimators of the total Kullback-Leibler discrepancy; the most used is the log-likelihood score. Statistical hypothesis testing is based on subsequent pairwise comparisons of log-likelihood scores of alternative models. Hypothesis testing allows one to derive a threshold below which the difference between two models is not significant and, therefore, the simpler models can be chosen.

Therefore, with statistical tests it is possible make an accurate choice among the models. The defect of this procedure is that it allows only a partial ordering of models, requiring a comparison between model pairs and, therefore, with a large number of alternatives it is necessary to make heuristic choices regarding the comparison strategy (such as choosing among the forward, backward and stepwise criteria, whose results may diverge). Furthermore, a probabilistic model must be assumed to hold, and this may not always be possible.

Criteria Based on scoring functions

A less structured approach has been developed in the field of information theory, giving rise to criteria based on score functions. These criteria give each model a score, which puts them into some kind of complete order. We have seen how the Kullback-Leibler discrepancy can be used to derive statistical tests to compare models. In many cases, however, a formal test cannot be derived. For this reason, it is important to develop scoring functions that attach a score to each model. The Kullback-Leibler discrepancy estimator is an example of such a scoring function that, for complex models, can be often be approximated asymptotically. A problem with the Kullback-Leibler score is that it depends on the complexity of a model as described, for instance, by the number of parameters. It is thus necessary to employ score functions that penalise model complexity.

The most important of such functions is the AIC (Akaike Information Criterion) (Akaike, 1974). From its definition, notice that the AIC score essentially penalises the loglikelihood score with a term that increases linearly with model complexity. The AIC criterion is based on the implicit assumption that q remains constant when the size of the sample increases. However this assumption is not always valid and therefore the AIC criterion does not lead to a consistent estimate of the dimension of the unknown model. An alternative, and consistent, scoring function is the BIC criterion (Bayesian Information Criterion), also called SBC, formulated by Schwarz (1978). As can be seen from its definition, the BIC differs from the AIC only in the second part, which now also depends on the sample size n. Compared to the AIC, when n increases the BIC favours simpler models. As n gets large, the first term (linear in n) will dominate the second term (logarithmic in n). This corresponds to the fact that, for a large n, the variance term in the mean squared error expression tends to be negligible. We also point out that, despite the superficial similarity between the AIC and the BIC, the first is usually justified by resorting to classical asymptotic arguments, while the second by appealing to the Bayesian framework.

To conclude, the scoring function criteria for selecting models are easy to calculate and lead to a total ordering of the models. From most statistical packages we can get the AIC and BIC scores for all the models considered. A further advantage of these criteria is that they can be used also to compare non-nested models and, more generally, models that do not belong to the same class (for instance a probabilistic neural network and a linear regression model).

However, the limit of these criteria is the lack of a threshold, as well the difficult interpretability of their measurement scale. In other words, it is not easy to determine if the difference between two models is significant or not, and how it compares to another difference. These criteria are indeed useful in a preliminary exploratory phase. To examine this criteria and to compare it with the previous ones see, for instance, Zucchini (2000), or Hand, Mannila, & Smyth (2001).

Bayesian Criteria

A possible “compromise” between the previous two criteria is the Bayesian criteria, which could be developed in a rather coherent way (see, e.g., Bernardo & Smith, 1994). It appears to combine the advantages of the two previous approaches: a coherent decision threshold and a complete ordering. One of the problems that may arise is connected to the absence of a general purpose software. For data mining works using Bayesian criteria the reader could see, for instance, Giudici (2001), Giudici & Castelo (2003) and Brooks et al. (2003).

Computational Criteria

The intensive wide spread use of computational methods has led to the development of computationally intensive model comparison criteria. These criteria are usually