A Heuristic Method of Model Choice for Nonlinear Regression

J. Ćwik and J. Koronacki

Institute of Computer Science, Polish Academy of Sciences,
ul. Ordona 21, 01-237 Warsaw, Poland
jc, korona@ipipan.waw.pl

Abstract. A heuristic method of model choice for a nonlinear regression problem on real line, based on the Equation Finder (EF) of Zembowicz and Żytkow (1992), is proposed and discussed. In our implementations of the EF we use a new, actually a three-stage, procedure for stabilizing model selection. First, a set of pseudosamples is obtained from the original sample by resampling in some way. Second, for each pseudosample, a family of acceptable models is found by a clustering-like algorithm performed on models with largest (adjusted) coefficients of determination. And third, the final selection is made from among the models which appear most often in the families obtained in the second stage.

1 Introduction and Outline of the Model Selection Procedure

Discovering equations or functional relationships from data in presence of random errors should be an inherent capability of every data mining and, more generally, knowledge discovery system. Such discoveries, in statistical terms, amount to choosing a model in a regression set-up. That is, given a random sample of pairs of variables, $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_i = f(x_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$, ε_i's are zero-mean random variables and f is an unknown function, the task is to estimate f by a member of a given family of nonlinear functions and to assess validity of the model thus obtained.

In this report, we present a new application of Equation Finder (EF), a discovery system of Zembowicz and Żytkow (see Zembowicz and Żytkow (1992) for a detailed description and discussion of the system and Moulet (1997) for the latter). The system finds "acceptable" models by means of a systematic search among polynomials of transformed data, when transformations – such as, e.g., logarithm, exponent and inverse – of both the independent and response variable are allowed. The family of all models to be considered is decided upon in advance by listing possible transformations of data and choosing the highest possible order of the polynomials. In the original version of EF, possible acceptance of a particular model is based on weighted least-squares (WLS) estimation of the model parameters and on the χ^2 test of fit.

Although EF uniquely determines the true model for samples from the same true model with sufficiently small error or large range of x values, it fails to provide sufficiently stable results for samples that carry larger error. As will be seen from simulation results in Section 3, for different samples from the same distribution the system provides results that may differ substantially from sample to sample. (Admittedly, this sort of instability is common to practically all the well-established methods of model choice in nonlinear regression.) It follows that some sort of stabilization of the original model selection procedure is needed, at least for regression problems with "larger errors".

In our implementations of the system, both the regression models and their parameters are provided by EF, the latter in accordance with the WLS method. But our choice of acceptable models (from those provided by EF) is different. It consists of three stages. First, a set of pseudosamples is obtained from the original sample by resampling combined with leave-many-out procedure: the original data are randomly permuted and the first pseudosample is obtained by leaving-out the first 10% of the data (for simplicity, we assume here that the sample size is divisible by 10); then the second 10% is left-out to obtain the second pseudosample, again of the size equal to 90% of the size of the original data, and so on, each time giving a pseudosample of the same size; once the last 10% of the data is left-out, the whole process, starting with permuting the original data, is repeated several times (e.g., if the sample size is 100, each permutation enables one to obtain 10 pseudosamples of size 90). Second, for each pseudosample, a family of acceptable models is found by a clustering-like algorithm performed on models with largest (adjusted) coefficients of determination. Looosely speaking, clustering rests on two-dimensional scaling based on distances between the models. And third, the final selection is made from among the models which appear most often in the families obtained in the second stage.

The rationale behind the method of choice of acceptable models is simple. Had we many independent sets of data, it would be reasonable to find a regression model for each set independently of the others, and then to aggregate the models obtained into one. Given just one set, we use it to produce many pseudosamples and proceed as if they were independent sets of data. In fact, our method of model selection borrows from stabilization ideas of Breiman (1996a) (see also Breiman (1996b)), with aggregation of multiple versions of a predictor[1] essentially made by a plurality vote: the final selection is made from among the models which appear most often in clusters of models obtained for subsequent pseudosamples. It is believed, and has been confirmed by simulations, that predictors which are relatively close to the true model underlying the data are likely to form clusters in a properly defined space of models.

In the next section, more details on the method will be given. Simulation results will be discussed in Sect. 3. Although we shall confine ourselves to discussing just one example, several more have been thoroughly investigated, leading to essentially the same conclusions. All in all, the simulations show that the method

[1] Borrowing from machine learning terminology, we use the terms regression model and predictor interchangeably.

proposed can be considered a useful tool for model selection in the nonlinear regression context.

In this report, we deal with discovering functions of one variable only, in accordance with the current capability of EF. In a separate report by the authors and J. Żytkow, possible extensions to discovering functions of several variables will be discussed.