Mathematical Modelling and Analysis. Ed. A.Buikis
Volume 12 Number 1, 2007, pages 81–90
 ISSN 1392-6292 print, ISSN 1648-3510 online

MODEL EVALUATION AND SELECTION IN MULTIPLE NONLINEAR REGRESSION ANALYSIS

G.JEKABSONS, J. LAVENDELS and V. SITIKOVS

Department of Informatics and Programming, Riga Technical University
Meza 1/3, LV1048, Riga, Latvia
E-mail: gintsj@egle.cs.rtu.lv; jurisl@egle.cs.rtu.lv

Received September 29, 2006; revised December 1, 2006; published online February 10, 2007

Abstract. The main problem in regression model selection is finding the best model that best fits the data, i.e. it does not neither overfit nor underfit. The aim of this work is to show one of possible ways to find adequate nonlinear regression models (parametric) of technical systems based on an heuristic search and analytical optimality evaluation approach by taking into consideration the computational power of modern computers.

Key words: Regression, approximation, model selection, heuristic search, model evaluation

1. Introduction

The main problem in regression model selection independently from application domain is finding the best model that best fits the data and does not neither overfit (the model is too complicated with too many free parameters) nor underfit (the model is too simple and therefore can't express the data sufficiently well). So the goal is to select a model that is the best trade-off between overfitting and underfitting. For this purpose we must consider more than just the model's error in our experimental data set which was used for estimation of parameter values. The models must be evaluated by using some kind of method that evaluates model's true predicting performance on yet unobserved data. Some of the most popular methods for model predicting performance evaluation are validation methods [7, 9] and information theoretic methods [1, 8, 12].

Due to a great computational power, now we have new possibilities for implementing methods, that demand expensive calculations to examine a big quantity of potentially optimal models. However the approach to evaluate all

possible models in a defined model space is impractical, since there still exist too many possible models to evaluate them all in acceptable time even with modern computers. This goal of this paper is to develop a heuristic approach for a selection of multiple (multidimensional) nonlinear regression model (parametric) by using heuristic state space search methods and information theoretic model evaluation methods. We use here the Bayesian Information Criterion (BIC) [12], which is already shown to be very effective in [8] and take into consideration the computational power of modern computers.

The paper is organized as follows. Section 2 describes one of possible approaches in regression model selection by using a heuristic search for the best model as a sum of components. Section 3 describes empirical comparisons of the heuristic search algorithms suitable for the approach. Section 4 summarizes the results of empirical comparisons and draws some conclusions about effectiveness of the search algorithms. Section 5 describes a practical application of this new approach. Conclusions are presented in section 6.

2. The Considered Approach

A regression model can be viewed as a sum of previously chosen known individual components (functions) of set $F\{f_i\}$:

$$\Phi = a_0 f_0 + a_1 f_1 + \cdots + a_{M-1} f_{M-1},$$

where a_i is model parameters, M is the number of used components, $f = f(X_1, X_2, \ldots, X_D)$ is function of independent variables, X_j is the j-th variable, D is the number of variables or data dimensions. In practical applications the number of such models can be very large.

The problem of regression model selection then can be formulated as follows: take a set of candidate components F and select a subset (not necessarily systematical) that performs best. This procedure can provide a better regression accuracy due to finite sample size effects, since model's irrelevant components may negatively affect the accuracy of regression [3, 6, 13]. In addition, reducing the number of components may help decrease the cost of acquiring data and might make the regression models easier to understand. Formally for solving the model selection problem the subset $F' \subseteq F$ should be found:

$$J(F') = \min_{F' \subseteq F} J(F''),$$

where $J(\cdot)$ stands for a chosen model evaluation criterion that should be minimised.

As the model evaluation criterion $J(\cdot)$ we are using information theoretic analytical model evaluation methods [1, 12]. It is shown in [8], that they are very effective for evaluation of regression models. The most straight-forward approach for searching the best model is to evaluate all possible models in model space and then to choose the best one (see Figure 1, where an example of model space (or state space) with four components is given). However such

approach is impractical, as typically there exist too many possible models to evaluate them all in acceptable time even with modern computers (especially in multiple nonlinear regression analysis).

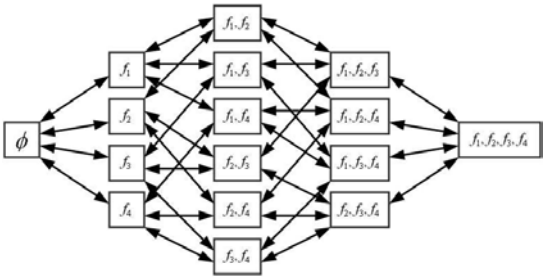

Figure 1. A small state space example.

A convenient paradigm for investigation of such problems is that of heuristic search [3, 6, 9] with each state in the search space specifying a possible model. In this case we can use heuristic search algorithms to traverse the space by adding and deleting components and select the best model.

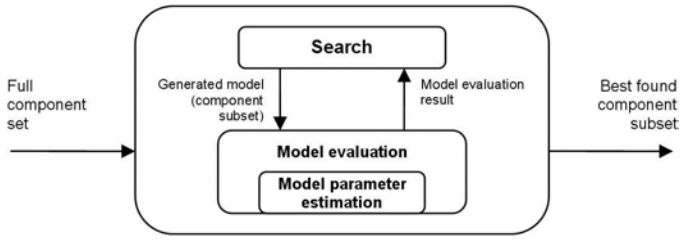

Figure 2. Regression model selection.

The process of regression model selection can be formed as shown in Figure 2. As input we have a full component set. Each time the search algorithm generates a new model to be evaluated, an evaluation algorithm estimates the parameters of the model and calculates the value of the criterion used. The calculated value is then used in further search process for guiding the search in the state space in the direction of possibly better models. When the search stops as output we have the best found component subset, i.e. the best found model.

3. Computational Experiments

In our performed experiments a special case was considered when all models are partial polynomials build of functions (components)

$$f_i(X) = \prod_{j=1}^{D} X_j^{r_{ij}},$$

where $r_{ij} = \{r_{i1}, r_{i2}, \ldots, r_{iD}\}$ is a vector of orders of features, $r_{ij} = 0, 1, \ldots, p$ is the order of the X_j variable, p is a previously chosen highest order. In addition the sum of all orders of all functions is less than or equal to p :

$$\sum_{j=1}^{D} r_{ij} \leq p.$$

Seven popular search algorithms together with the Bayesian Information Criterion (BIC) [12] (already shown to be very effective in [8]) were considered for empirical evaluation of effectiveness from the aspect of both optimality of the results and necessary computing resources. In addition they were compared with full polynomial models. For the experiments we used our software that was developed by using Delphi Object Pascal. It implements empirical experiments with various multidimensional data, search algorithms, and model evaluation criteria.

We have compared the following algorithms (the first four of them are sequential and the last three are stochastic):

1. Sequential Forward Selection (SFS),
2. Plus l Take Away r Selection (PTA),
3. Sequential Floating Forward Selection (SFFS),
4. Hill Climbing (HC),
5. Random-Restart Hill Climbing (RRHC),
6. Random-Mutation Hill Climbing (RMHC),
7. classic Genetic Algorithm (GA).

These and other heuristic search algorithms are discussed in [3, 6, 9, 11]. We have used the following true error rate estimation criteria [2, 7, 8, 10]:

1. Test data set Average Absolute Error, AAE.
2. Test data set Relative Root Mean Square Error (RRMSE)

$$RRMSE = \sqrt{\frac{MSE}{Variance}} = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(\Phi(x_i) - y_i)^2\right) \Big/ \left(\frac{1}{N}\sum_{i=1}^{N}(\overline{y} - y_i)^2\right)},$$

where x_i are the input variables of the i-th training sample, y_i is the response variable of the i-th training sample, $\overline{y}$ is the mean value of response values in all samples, N is the number of samples, MSE stands for the Mean Square Error.

In all experiments for each algorithm the search time, a number of found model's components, the value of found model's evaluation criterion and values of found model's true error estimations were recorded. The results are shown in tables where four best search results and one result for the best full polynomial are presented. We note that all stochastic algorithms were started in randomly chosen states in search space.

3.1. Experiment with Hwang's 4th function with noise

In this experiment a data set was generated by using the Hwang 4th function (see [5]) which is proposed for testing of regression methods. The values of both features X_1 and X_2 are uniformly distributed over $[0, 1]$. The following parameters of the problem are used in experiments: the number of variables is equal to 2, the maximum polynomial order is 8, the number of all components used to generate models is 45. The learning data set was formed of 100 randomly chosen examples from all the data and the test data set was formed of 10000 randomly chosen examples from all the left data. The results are presented in Table 1.

Table 1. Results of the experiment.

	Full5	SFFS	RRHC	RMHC	GA
Time	$<$ 1s	0.1 s	12 s	2.5 s	1.4 s
# of comp.	21	8	12	14	12
BIC	-168.0	-97.63	-254.5	-251.3	**-257.6**
AAE	0.3215	0.6289	**0.2107**	0.2203	0.2132
RRMSE, %	43.89	97.98	**25.70**	27.18	26.03

In this experiment even with such a small number of used components all sequential algorithms found only local minimum values. The stochastic algorithms were more effective, they effectively avoided these local minimums. All of the found models have much smaller AAE and RRMSE error values than the best full polynomial.

3.2. Experiment with Friedman's function

This is an artificial data set from [4]. The examples are generated using the following method. First we generate the values of 10 variables, $X_1, X_2, \ldots, X_{10}$ uniformly distributed over $[0, 1]$. Then we obtain the values of the target feature by using the equation ($X_6 \ldots X_{10}$ are not used):

$$P_{Fried}(X) = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon,$$

where ε is a normal distribution noise with the mean value equal to 0 and the dispersion equal to 1.

The number of variables is equal to 10, the maximum polynomial order used is 3, the number of all components used to generate models is 286. The learning data set was formed of 1000 and 400 randomly chosen examples from all the data, the test data set was formed of 10000 randomly chosen examples from all the left data. The results are presented in Tables 2 and 3.

The presented results of experiments with the Friedman function have shown that by reducing the number of examples in a learning data set the

Table 2. Results of the experiment with 1000 data points.

	Full3	SFS	PTA	SFFS	HC
Time	7 s	36 s	59 s	40 s	71 s
# of comp.	286	17	15	12	12
BIC	1789	228.0	219.3	**209.9**	**209.9**
AAE	0.9817	0.8346	0.8371	**0.8337**	**0.8337**
RRMSE, %	24.87	20.98	21.05	**20.97**	**20.97**

Table 3. Results of the experiment with 400 data points.

	Full2	SFS	PTA	SFFS	HC
Time	< 1 s	14 s	17 s	13 s	22 s
# of comp.	66	18	13	12	12
BIC	693.1	118.3	98.04	**88.86**	**88.86**
AAE	1.469	0.8468	0.8669	**0.8418**	**0.8418**
RRMSE, %	37.73	21.30	21.80	**21.18**	**21.18**

accuracy of full polynomials decreases much faster than the accuracy of partial polynomials. In the experiments with 1000 data points the order of the best full polynomial was 3, but in experiments with 400 data points it was equal to 2. However the best found partial polynomial model was the same in both experiments. It was found by SFFS and HC algorithms and it was built entirely of the five features that are used in Friedman's function equation.

4. Summary of Experimental Results

Obtained results of the performed experiments prove that the considered approach is effective in multiple nonlinear regression analysis model selection.

In experiments with a relatively small number of components the best results were obtained by using RRHC and GA algorithms. However when a time consumption is taken into consideration the best trade-off is given by RMHC. Its time consumption is much smaller than CPU time of RRHC, and the obtained models are almost of the same quality.

When the number of used components is relatively big (approx. 100 and more components) the best results were obtained by using sequential algorithms. Apparently the state space is too big for stochastic algorithms to be effective with so small number of iterations allowed. By increasing the number of iterations it is possible to find better results, however in such a case computing resources needed for the search are greatly increased. In experiments with relatively big number of components the best results were obtained by using the HC algorithm. However the best trade-off between obtained model's evaluation and time consumption was SFFS.

Our overall conclusion about the search algorithms is that effectiveness of the algorithms depends on the addressed problem, i.e. the greater the number

of components is used for model generation, the more effective are sequential search algorithms. When the number of components for model generation grows up, the probability of a sequential algorithm to get stuck in a local minimum reduces. However computing resources needed for the search for stochastic algorithm quickly increases. When the number of components is relatively small sequential algorithms get stuck in local minimums very often. We note that in such cases there are relatively many local minimums in the state space. In contrast in such cases stochastic algorithms perform better by avoiding local minimums (GA, RMHC) or restarting the search from different states (RRHC, RMHC).

Also we conclude that the biggest benefit from the considered approach is obtained when the available data is relatively small, which is a frequent phenomena in real-world practical applications.

5. A Practical Application

The obtained results can be used for a regression model selection in empirical data of various origins for developing optimal technological solutions. For example, we did experiments with regression model acquisition of aircraft shell's behavior during pre-buckling and post-buckling phases (see Fig. 3) which allows to determine at what amount of the load the construction loses its stability and collapses [7]. Previously, full polynomials of 2nd and 3rd order were used in a similar analysis.

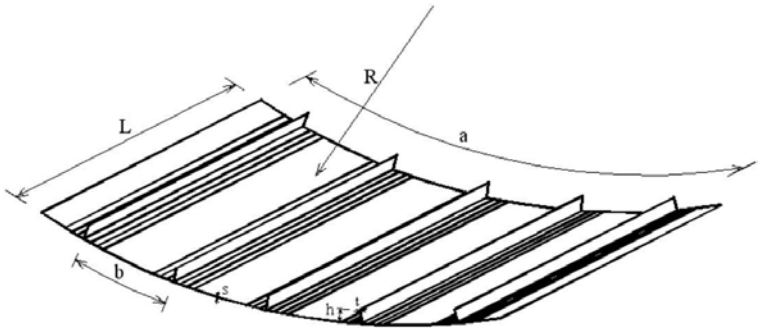

Figure 3. Dimensions of a shell.

Four design variables of the construction are described as follows: the number of stiffener $x_1 = n(4-6)$, the height of stiffener $x_2 = h(15-20)$, the internal radius $x_3 = R(800-1600)$ and the panel length $x_4 = L(600-720)$. The maximum load value representing the numerical collapse load is denoted by P_{co}. The principal goal of the investigation is to build surrogate models, where P_{co} is the function of design parameters. The experimental design was optimized according to the Mean Square Error (MSE) criterion. For each

Table 4. Results of the approximations.

Method	BIC	AAE	RRMSE	# of comp.
Full2	303.5	5.85	12.04 %	15
Full3	332.4	5.88	12.42 %	35
SFFS	375.6	6.85	14.58 % 5	
RRHC	**267.7**	**4.69**	**9.87** %	9

number of stiffeners the three-dimensional design space with 27 sample points was elaborated, in total we did $3 \times 27 = 81$ experiments.

Table 4 presents the results of approximation of P_{co}. The empirical data was approximated with full polynomials of 2nd and 3rd order (Full2 and Full3) as well as with partial polynomials of 3rd order. In the search for the best partial polynomial model two of the most promising search algorithms were used, i.e. SFFS and RRHC. The results show that the best model is found by RRHC algorithm, its both BIC values and the AAE and RRMSE values are better than of any other considered model. SFFS algorithm, as we can see, apparently was stuck in a local minimum at some rather early search step and therefore its best found model is even worse than the models obtained by Full2 or Full3.

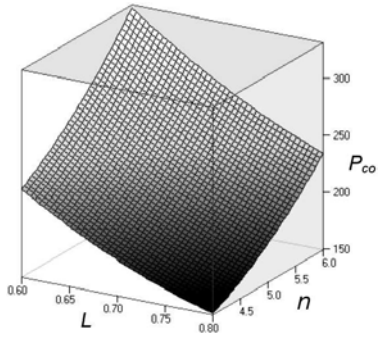

Figure 4. The surface plot of the best found model.

In such a way by using the RRHC algorithm and BIC criterion we found a model that is more accurate and more simple (fewer components) than the full polynomial models that were used before. In Figure 4 the surface plot of the best model obtained by the RRHC algorithm is shown, when $h = 0.02$ and $R = 0.9385$ are fixed.

6. Conclusions

This paper reflects a research goal to develop a heuristic approach for multiple nonlinear regression model selection by using heuristic state space search

methods and information theoretic model evaluation methods and by taking into consideration the computational power of modern computers.

The obtained results of the performed experiments prove that heuristic search methods and analytical model evaluation criteria are effective tools for model selection in a multiple nonlinear regression analysis.

The results of the theoretical research are implemented in software that can be used for regression model selection in empirical multidimensional data of various origins by developing optimal technological solutions. The developed software with implemented various search algorithms and model evaluation criteria is already used for modeling applications at Institute of Materials and Structures, Riga Technical University. In these experiments the obtained models are almost always more effective than previously used. The developed software is effective and competitive tool for solving practical regression model selection problems.

7. Acknowledgement

This work has been partly supported by the European Social Fund within the National Program "Support for the carrying out doctoral study program's and post-doctoral researchers" project "Support for the development of doctoral studies at Riga Technical University".

References

[1] H. Akaike. A new look at the statistical model identification. In: *IEEE Transactions on Automatic Control*, volume 19, 716–723, 1974.

[2] J. Auzins, K. Kalnins and R. Rikards. Sequential design of experiments for metamodeling and optimization. In: *Proceedings of the 6th World Congresses of Structural and Multidisciplinary Optimization*, Rio de Janeiro, Brazil, 2005.

[3] M. Dash and H.Liu. Feature selection for classification. In: *Intelligent Data Analysis. An International Journal*, volume 1. Elsevier, 131–156, 1997.

[4] DELVE. Data for evaluating learning in valid experiments. . http://www.cs.toronto.edu/ delve/

[5] J.N. Hwang, S.R.Lay, M.Maechler, R.D.Martin and J.Schimert. Regression modeling in back-propagation and projection pursuit learning. In: *IEEE Transactions on Neural Networks*, volume 5, 342–353, 1994.

[6] G. Jekabsons. Reducing hypothesis complexity in multiple regression. In: *Computer Science, Scientific Proceedings of Riga Technical University*, volume 22. RTU, Riga, 50–62, 2005.

[7] G. Jekabsons and K. Kalnins. Selection of construction behavior metamodel employing heuristic state space search. In: *Scientific Proceedings of Riga Technical University, Computer Science*, volume 22. RTU, Riga, 266–276, 2005.

[8] G. Jekabsons and J. Lavendels. Evaluation of model selection criteria in multiple nonlinear regression analysis. In: *Computer Science, Scientific Proceedings of Riga Technical University*, volume 23. RTU, Riga, 67–81, 2005.

[9] G. Jekabsons and J. Lavendels. A heuristic approach of model selection in multiple nonlinear regression analysis. In: *Proceedings of the IADIS International Conference, Applied Computing 2006, Mondragon unibertsitatea, Spain, San Sebastian*, 524–527, 2006.

[10] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: C.S. Mellish, Morgan Kaufmann(Eds.) *Proceedings of IJCAI-95*, 1137–1143, 1995.

[11] A. Roverato and S.Paterlini. Technological modeling for graphical models: an approach based on genetic algorithms. In: *Computational Statistics & Data Analysis. Applications of Optimization Heuristics to Estimation and Modelling Problems*, volume 47, 323–337, 2004.

[12] G. Schwarz. Estimating the dimension of a model. In: *Annals of Statistics*, volume 6, 461–464, 1978.

[13] V. Vapnik. *Statistical Learning Theory.* John Wiley, USA, 1998.