A HEURISTIC APPROACH OF MODEL SELECTION IN MULTIPLE NONLINEAR REGRESSION ANALYSIS

Gints Jekabsons
Department of Informatics and Programming, Riga Technical University
Meza 1/3, LV1048, Riga, Latvia

Juris Lavendels
Department of Informatics and Programming, Riga Technical University
Meza 1/3, LV1048, Riga, Latvia

ABSTRACT

This paper reflects a research goal of which is to develop heuristic approach for multiple nonlinear regression analysis model selection.

From sixteen heuristic search algorithms suitable for multiple nonlinear regression analysis eight most popular algorithms were considered. All of the algorithms were classified and empirically evaluated from the aspect of both necessary computing resources and optimality of the results.

The theoretical results of the research are implemented in software, which was used for approbation of the described approach in construction behavior modeling applications at Institute of Materials and Structures, Riga Technical University. Models obtained were more effective than previously used. Developed software is effective and competitive tool for solving practical regression problems.

KEYWORDS

Regression, approximation, model selection, heuristic search methods

1. INTRODUCTION

In empirical risk minimization there can be differentiated two problem formulations: classification and regression [Vapnik, 1998]. In both fields two following steps can be marked out:

- model selection (still actual problem in nowadays);
- calculation of model's parameters (the greatest part of the methods for regression is developed already some centuries ago).

Regression model can be viewed as a sum of individual components of set $F = \{F_i\}$:

$$\Phi = a_0 F_0 + a_1 F_1 + \ldots + a_{M-1} F_{M-1}, \tag{1}$$

where a_i is model parameters. In practical applications the number of possible models can be a very significant figure.

Model selection and calculation of parameters can be done in succession:

- if characteristics of functional relationships being investigated are known and model is selected from the aspect of these characteristics;
- if the characteristics are unknown and it is assumed that the process can be described by polynomial of certain degree, Harmonic series etc.

Model selection and calculation of parameters can be done in turns:

- by gradually increasing model's complexity without recalculation of already calculated parameters (Chebishev's method and its variations);

- by increasing or decreasing number of components in model which is built by certain rules (change of full polynomial's degree and recalculation of all its parameters; the same with Harmonic series).

Considered approach:

- deems that optimal regression model can also contain elements of set *F* unsystematically;
- does model selection based on analytical model optimality evaluation methods [Akaike, 1974; Schwarz, 1978] and heuristic search methods [Dash and Liu, 1997; Jekabsons et al^1, 2005].

2. REGRESSION MODEL SELECTION

When searching for the optimal model overfitting is the major problem in both classification and regression [Vapnik, 1998]. In both these fields we must find a way how to generate best model and how to evaluate its optimality.

The problem of model selection is to take a set of candidate components and select a subset that performs best. This procedure can provide better regression accuracy due to finite sample size effects – irrelevant components may negatively affect the accuracy of regression [Dash and Liu, 1997; Jekabsons et al^1, 2005; Vapnik, 1998]. In addition, reducing the number of components may help decrease the cost of acquiring data and might make the regression models easier to understand.

When evaluating model optimality the most straight-forward way is to simply choose a model that has the smallest error in learn data set. However the selection process is complicated by the fact that a model with many free parameters is more elastic than a simple model with less free parameters and therefore it can fit to the data better (possibly causing overfitting). So the point is to select a model that is the best trade-off between fitting to data well and fitting to it too well – overfitting it. For this purpose for example validation methods [Kohavi, 1995] or analytical methods [Akaike, 1974; Schwarz, 1978] can be used.

When generating the optimal model the most straight-forward way is to evaluate all possible models and then to choose the best one. However such approach is impractical, as there exists too many possible models to evaluate them all in acceptable time. A convenient paradigm for viewing such problems is that of heuristic search, with each state in the search space specifying a possible model [Dash and Liu, 1997; Jekabsons et al^1, 2005]. In this case we can use heuristic search algorithms to traverse the space and select a model that is optimal or close to optimal.

3. EVALUATION OF THE ALGORITHMS

To evaluate effectiveness of most popular heuristic search algorithms suitable for multiple nonlinear regression analysis empirical approach was used. For this purpose software for empirical experiments with various multidimensional data, search algorithms, and model evaluation criterions was developed.

By using the software a number of experiments were performed in which a special case was discussed when all considered models are partial polynomials build of functions

$$F_i = \prod_{j=1}^{D} X_j^{r_{ij}}, \tag{2}$$

where $r_i = \{r_{i1}, r_{i2}, \ldots, r_{iD}\}$ is vector of orders of features; $r_{ij} = 0,1,\ldots,p$ is order of the X_j feature; *p* is previously chosen maximal order. In addition the sum of all orders for all functions is lower or equal with *p*:

$$\sum_{j=1}^{D} r_{ij} \le p. \tag{3}$$

Eight heuristic search algorithms were evaluated from the aspect of both necessary computing resources and optimality of the results. The first five of them are sequential and the last three are stochastic:

- Sequential Forward Selection (SFS);
- Sequential Backward Selection (SBS);
- Plus *l* Take Away *r* Selection (PTA);
- Sequential Floating Forward Selection (SFFS);

- Hill Climbing (HC);
- Random-Restart Hill Climbing (RRHC);
- Random-Mutation Hill Climbing (RMHC);
- simple classic Genetic Algorithm (GA).

These and other heuristic search algorithms are discussed in [Dash and Liu, 1997; Jekabsons et al^1, 2005]. Table 1 shows classification of considered algorithms in view of the discussed problem by using heuristic search algorithm classification in [Dash and Liu, 1997].

Table 1. Classification of the evaluated algorithms

Sequential algorithms			**Stochastic algorithms**	
Forward generation	Backward generation	Combined generation	Pure random generation	Probabilistic generation
SFS SFFS	SBS	PTA HC	RMHC RRHC	GA

Model evaluation criterion used in experiments: Bayesian Information Criterion (BIC) [Schwarz, 1978].

Model true error rate estimation criterions used in experiments [Jekabsons et al^2, 2005; Kohavi, 1995]:

- Test data set root mean square error percent;
- Test data set average absolute error.

In all experiments each algorithm's search time, number of found model's functions, value of found model's BIC criterion and values of found model's true error rate estimations were recorded. For comparison full polynomial model evaluations were also recorded.

4. SUMMARY OF EXPERIMENTAL RESULTS

Obtained results of the performed experiments approve that heuristic search algorithms are effective in multiple nonlinear regression analysis model selection. And following conclusions can be drawn: the greater the number of functions used for model generation, the more effective are sequential search algorithms. When the number of functions is smaller sequential algorithms get stuck in local minimums much often. That is probably because in such cases there are relatively more local minimums in the state space. For comparison in such cases stochastic algorithms perform better by avoiding local minimums (GA, RMHC) or restarting the search from different starting states (RRHC, RMHC). When the number of functions is bigger better 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's possible to find better results however in such case computing resources needed for the search would greatly increase.

In experiments with relatively big number of functions best results were obtained by using HC algorithm. However the best trade-off between obtained model's evaluation and time consumption of the algorithm was SFFS.

In experiments with relatively small number of functions best results were obtained by using RRHC and GA algorithms. However when also time consumption is taken into consideration best trade-off is RMHC. Its time consumption is much smaller than RRHC's but obtained models are almost as good.

Overall conclusion is that efficiency of the algorithms depends on addressed problem – the greater the number of functions used for model generation, the more effective are sequential search algorithms. When the number of functions used for model generation grows, probability of sequential algorithm to get stuck in local minimum reduces. However computing resources needed for the search for stochastic algorithm quickly increases.

Also concluded was that the biggest benefit from partial polynomials is when the available data for analysis is relatively small.

5. CONCLUSIONS

This paper reflects a part of the research goal of which is to develop heuristic approach for multiple nonlinear regression problem solving.

The main results:

- The effectiveness of heuristic search methods for model selection in multiple nonlinear regression analysis is shown.
- By basing on the results of research software for empirical experiments with various data, search algorithms, and model evaluation criterions was developed.
- Empirical evaluation of effectiveness of the most popular heuristic algorithms from the aspect of both necessary computing resources and optimality of the results was performed. It was observed that the biggest benefit from partial polynomials is when the available data for analysis is small.
- In performed experiments it was found that heuristic search algorithm's efficiency depends on number of functions used for model generation. Specific recommendations for what kinds of algorithms are most effective in each type of problem are given.
- The developed software was used for approbation of the described approach in construction behavior modeling applications at Institute of Materials and Structures, Riga Technical University. Obtained models were better than previously used. The software is competitive tool for solving of practical regression problems.

Further experimental work in multiple nonlinear regression analysis model selection may consist of considering other types of functions: with negative order, exponential etc. Orthogonal polynomials may also be discussed.

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

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, pp. 716-723.

Dash, M. and Liu, H., 1997. Feature Selection for Classification. *Intelligent Data Analysis - An International Journal, Elsevier*, Vol. 1, No. 3, pp. 131 – 156.

Jekabsons, G., et al^1, 2005. Reducing hypothesis complexity in multiple regression. *Computer Science, Scientific Proceedings of Riga Technical University*, Vol. 22. RTU, Riga, pp. 50-62.

Jekabsons, G., et al^2, 2005. Metamodels for the optimum design of corrugate load-bearing profile plates. *Architecture and Construction Science, Scientific Proceedings of Riga Technical University*, Vol. 2, No. 6. RTU, Riga, pp. 136-145.

Kohavi, R. and Mellish, C. S. (ed.), 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of IJCAI-95*, Morgan Kaufmann. pp. 1137-1143.

Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics*, Vol 6, pp. 461-464.

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