INTELLIGENT SYSTEM DESIGN FOR KNOWLEDGE STRUCTURE MODELS FROM OBSERVED DATA

VLADIMIR STEPASHKO[1], TATIANA ZVORYGINA[2]

International Research and Training UNESCO Center of Information Technologies and Systems, Ukraine, Kiev, Academic Glushkov Prospect, 40, 03680.
[1] Professor, Head of Department, E-mail: step@g.com.ua
[2]Master of Computer Science, Post-graduate Student, E-mail: zvortf@ua.fm

Abstract: The task of modeling from data observed is considered as a sequence of stages, or subtasks, of solutions choosing. It is shown that at each stage there is some finite subset of possible solutions depending from those accepted on the previous stages. Each of accepted solutions, therefore, restricts subsets of possible solutions at all consequent stages. It allows to organize an "intellectual interlayer" between a user who needs to model something and a modeling software. The intellectuality level of a system of such kind is determined by implementation of knowledge (both theoretical and practical) about a process being modeled as well as methods of modeling and by minimization of requirements to skills of a user.

Keywords: knowledge structuring, modeling, data observed, decision making, dialog shell, intellectual system.

1. INTRODUCTION

There are known modeling tasks to which one or another method of analysis is being applied in practice, based on which the experts consider to have as the most adequate method for the given purposes. However, for the majority of real problems it is not possible to specify in advance the exact line-up of operations, as there is no a priori information on a plant or process beeing modeled. Most importantly, people who need, for example, to predict some economic or ecological indexes are not experts in the modeling field.

The modeling software existing in the market, for example "Statistics", have one but very essential, in our opinion, shortcoming, they are mainly oriented on users possessing high qualification in the modeling field. An expert in the field can only tell which a method of parameter estimation, or a generator of model structures, or a criterion of model selection should be preferred. An economist or ecologist is forced to choose methods "at random" and then manually check the obtained models with respect to their correspondence to the purposes of research.

Therefore there is a necessity for creation of some "intellectual interlayer" between a user who needs to model something and modern computational software. This "intellectual interlayer" should be capable of advising the user in a dialog mode which method may be better.

It is our aim to create such an intellectual interlayer. The first problem we are faced with is one of classification of the available knowledge in the modeling field. The main part of such expertise exists in the form of practical experience on applying one or another method to specific problems as well as the limitations on their application.

By a method of modeling we mean a set of operations with a given data sample allowing one to build a mathematical relationship between the output variable and the input variables.

2. KNOWLEDGE CLASSIFICAITON

Let's assume that the data sample do not contain missing values of variables and are prepared for handling. There are quite well defined mathematical methods for data pre-processing [Duke, Samoilenko, 2001], so we shall not discuss them here.

We claim that each method of modeling contains, in explicit or implicit form, such key elements as a model class, an external criterion of model selection, a generator of model structures and a method of model parameters estimation. For example, it is easy to see that in the classical regression analysis, polynomial functions form the class of used models, inclusion or/and exclusion method is to be a generator of model structures, the

 ISBN 3-936150-25-7

least squares method (LSM) is used as the method of the model parameter estimation, and the Fisher criterion is one for model selection. If we try to identify similar components in the Akaike method, we shall get, accordingly: ARIMA is the class of used models, embedded structures are to be a generator of model structures, the Yool-Walker method (YWM) is used as the method of the model parameter estimation, and the Akaike criterion is one for model selection.

It was also appointed that a choice of a modeling method is affected by such circumstances as the purpose of investigations and the type of plant being modeled.

3. SUB-PROBLEMS

Therefore, in solving the problem of a relevant modeling method choice we define the following sub-problems to be solved sequentially:

1. Choice of the modeling purpose (approximation, interpolation, extrapolation, trend definition, prediction, construction of input-output model etc.)

2. Definition of the plant type (linear static, nonlinear static, linear time series, nonlinear time series, linear dynamic, and nonlinear dynamic)

3. Definition of process stationarity (stationary, with an increasing trend, with a decreasing trend, with an oscillatory trend, with the mixed trend)

4. Choice of a model class (linear regression, autoregression, autoregression with trend, harmonic, logarithmic, polynomial or exponential functions of time, difference equations etc.)

5. Choice of external criterion of model selection (Akaike criterion, “jack-knife”, C_p-statistics of Mallows, unbiasedness and/or regularity criterion, Fisher criterion etc.)6. Choice of a parameter estimation method (LSM, LMM, ridge regression etc.)

7. Choice of structure generation method (a given structure, embedded structures, inclusion, exclusion, inclusion-exclusion, exhaustive search, branches and bounds, combinatorial, combinatorial-selective, multilayer (GMDH)).

8. The obtained model validation (Fisher statistics, precision on control sample, etc.)

We claim that for the solution of problem of choosing a modeling method addressing to this set of subproblems is necessary and sufficient.

This order for solving the problem is well motivated. It was observed that the solution of the first subtask leads to the essential diminishing of the set of subsequent solutions. This happens because the decision making on each of stages introduces implicit limitations on application of these or other techniques the need in which to be decided at the consequent stages.

4. SOLUTION TECHNIQUES

After having determined the order of solving the subtasks, we have encountered the problem of finding such mathematical and dialogue procedures that would facilitate solving the formulated subtasks by an inexperienced (in the modeling field) user. To solve this problem we have used such modern techniques as Data mining [Duke, Samoilenko, 2001; Stepashko, 1991] and knowledge elicitation [Gavrilova, Khoroshevsky, 2000]. By combining these techniques with the dialogue interview of the user about the modeled plant, we have found a solution tree for the problem of observed data modeling. Even to a user at the first time facing the problem of modeling, this procedure allows to construct a more or less acceptable mathematical model.

It is better to give an example to illustrate our findings (see figure 1)

The four stages of decision making chosen for illustration are: definition of a process type, linearity, stationarity, and choice of the model class.

The first stage is choice of process type from the three indicated alternatives. If the user is not able to make a choice on his own, the dialogue and control tools help him. The principle of organization of the dialogue at the stage of deciding the process stationarity is described in detail in [Stepashko, Zvorygina, 2001].

It is easy to see from figure 1 that an investigator has very large set of model classes for choosing at the fourth stage (in the figure, we show only an incomplete class of models). If one attempts to solve the problem by the "brute force", each of possible models needs to be tested, what, in practice, is a sufficiently labour-consuming process. However, if one takes the advantage of decomposing the problem into the proposed subtasks, then, depending on the decision accepted at the first three stages, the set of allowed solutions will be significantly narrowed. For instance, if at

the early stages the decision is taken that the plant is linear and static, a unique solution is to use the linear regression model. If, however, the plant is linear and dynamic, the difference equation models can only be applied. In the case of linear time series without a trend, the model of autoregression is applied. If the trend is increasing or decreasing, it is possible to use a model of a linear trend. And in the case of an oscillatory trend, the choice is limited to a harmonic series or autoregression.

If at the first stage the decision is taken that the process is a "time series", at the second stage that it is non-linear, and at the third stage that it has an increasing trend (non-linear time series), then, at the stage of choosing the basis functions, the intellectual system will recommend the investigator to model the process in auto regression with trend model class. The user will also be given a choice of exponential and logarithmic functions of time. On the other hand, a whole class of models will be eliminated, such as linear regression model, or a model with a linear trend, etc. The considered variant is shown in figure 1 by shadow.

Note that the proposed organization of the intellectual envelope considerably simplifies the checking of correctness and consistency of the accepted decisions.

5. DISCUSSIONS AND CONCLUSIONS

The proposed principle of the intellectual envelope for a computer-aided data modeling system will have the following major advantages: interactive component at all stages of modeling; minimization of requirements to the user qualification; active utilizing the user knowledge base; constant monitoring and testing the accepted decisions; visualisation of the process of problem solving and contextness in perceiving the information; training the user during interaction with the system.

REFERENCES

Duke V., Samoilenko A., 2001 “Data mining”, St. Petersburg, Russia (in Russian)

Stepashko V., 1991, “On Expert Knowledge Structuring Task in the field of modeling from empirical data”, *Cybernetics and Computer Engineering, Iss. 92,* Kiev, Ukraine (in Russian)

Gavrilova T., Khoroshevsky V., 2000, “Knowledge Bases of Intellectual Systems”, St. Petersburg, Russia (in Russian)

Stepashko V., Zvorygina T., 2001, “On Design of a DSS Dialog Shell for Observation Data Modeling”, *Modeling and Control of State of Ecologic and Economic Systems of a Region,* Kiev, Ukraine (in Russian)

Volodymyr Stepashko

Doctor of Sciences (1994), Head of Department on “Information Technologies of Inductive Modeling” of International UNESCO Center of NASU, Kyiv.
Professor of the Joint Chair of the International UNESCO Center and of the National University “Kyiv Polytechnic Institute” for preparing of Masters of Sciences in specialty “Intellectual Systems of Decision Making”, special course “Inductive Approach to Complex Systems Modeling”.
Field of interests: data-based modeling of complex systems, system identification, control systems, intellectual knowledge-based systems, decision making.

Tetyana Zvorygina

Graduated from Kiev State University in 1992, specialty physicist. At 1999 entered to International Research and Training UNESCO Center of Information Technologies and Systems where get a Master degree in specialty intellectual system of decision making. Field of interest - intellectual system of knowledge extraction, data based modeling, decision making procedures.

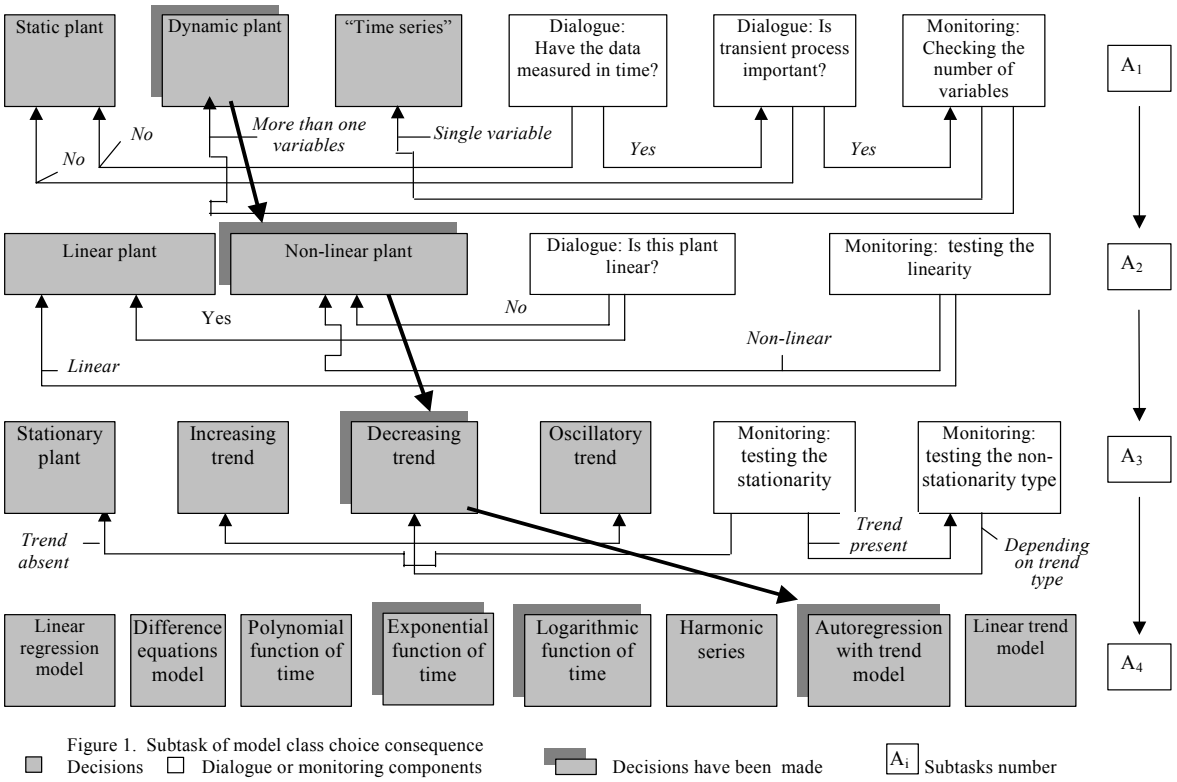

Figure-1: The 4 stages of decision making