

Formalization of Information Storing Structures in the Tasks of Inductive Modeling

Nataliya Shcherbakova, Volodymyr Stepashko

*International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine,
Glushkov ave., 40, Kyiv, 03680, Ukraine*

nataliya.shcherbakova@gmail.com, stepashko@irtc.org.ua

Abstract. *When solving real tasks of model construction from statistical data, the question arises regarding storage of and providing effective access to the information. To solve such a problem, an integrated environment of information storage is developed. Architecture of the environment is offered giving the possibilities to manipulate present information freely due to using relational database which contains only metadata and storage of input statistical data and results of calculations. The main question arising when developing the environment is formalization of information storing structures within the environment. Namely, a format of data storing and presentation of data structures in the storage. During the processing of input data, regardless of a modeling method, data come from different sources (often data contain omissions and atypically small/big values) and should be reduced to an unified form. On the other hand, there is a question of storing the output data, namely the structure and parameters of models, evaluation of reliability and accuracy, graphics etc. These questions are analyzed in the paper.*

Keywords

Integrated environment, handling and storing of
information, inductive modeling, GMDH

1 Introduction

Algorithms of inductive modeling are effective for use in solving practical problems of modeling economical, environmental, technological and other complex processes and systems [1, 2]. The problem of means of storing and handling of research data and using their results is of current importance. This article discusses the possible difficulties that arise when developing an integrated environment of handling and storing information in the tasks of inductive modeling. Attention focuses on the generalizing and structuring of main formats incoming data, development of standardized formats of storing calculations results and supporting information.

2 Architecture of the system

In [3] an architecture of integrated environment of handling and storing information in the tasks of inductive modeling was proposed, which provides freely manipulate available information through building layout environment which consists of relational database [4,5,6] containing only meta-data and XML storage, in which input data and results of calculations are stored [7].

Proposed layout of the environment is intended for solving problems of storage of input data and results. Designed architecture of environment of handling and storing information in the tasks of inductive modeling (Fig. 1) provides the opportunity to develop a software. Modular system architecture makes it possible to expand its functionality.

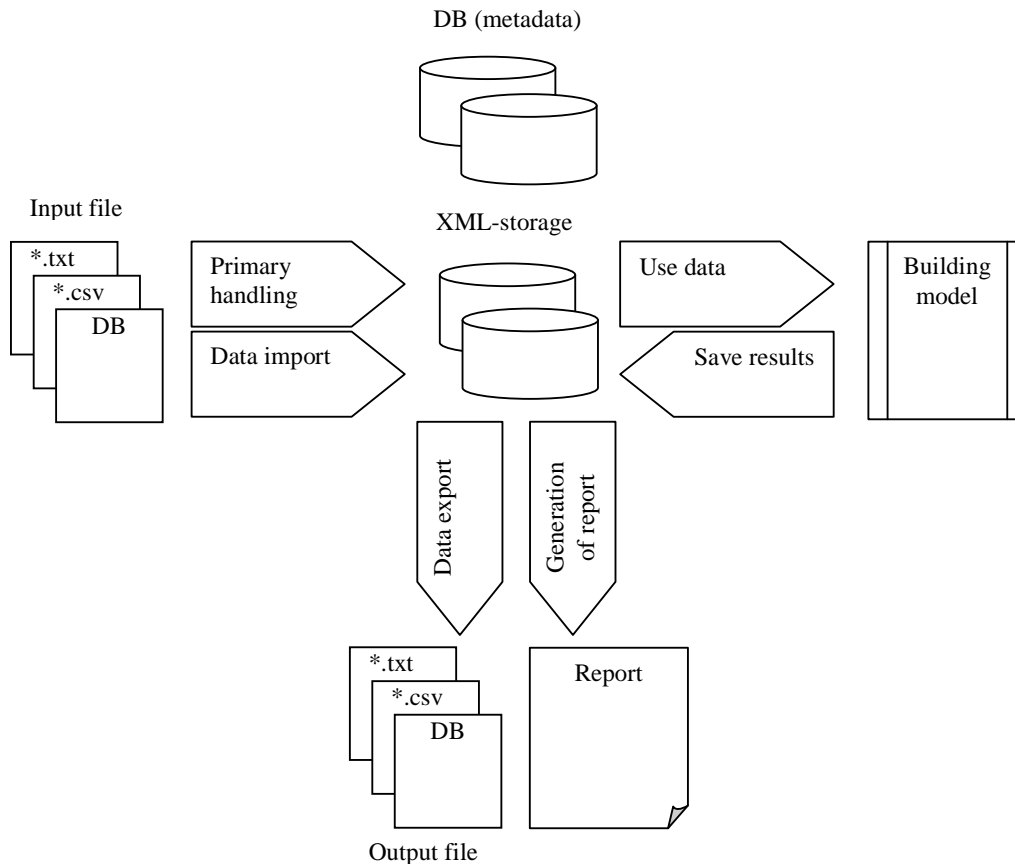


Fig. 1. Architecture of an integrated environment of information handling and storing.

The main requirements to the system is the ability to import (including primary processing) and export data, storing and handling existing information, storing output data with all information of calculation results, generate reports on the results. It should be noted that results of calculations are stored in the system in a standardized form that will allow generating strictly formal reports on the results of calculations. Open question remains on generalization and structuring of main formats of input data and development of standardized formats of storing the results of calculations.

Let us consider in more detail what information one needs to save in the system. Firstly, as already discussed above, these are input statistical data given to a single format and processed data with eliminated omissions and/or atypical values etc. Secondly there are basic functions, generated models, estimates of the parameters, criteria of quality models and best models. All the information is better to store in the XML-storage. Auxiliary information such as data on the user, date and time using of files etc. it is better to store in relational database.

3 Primary formalization of information storage structures in the tasks of inductive modeling

Solving real problems of models building using GMDH algorithms with help of the developed environment can be divided into the following stages:

- import input data from different sources;
- initial processing of input data;
- choose of the class of models;

- choose of an algorithm structures generator;
- choose of a method of parameters estimation;
- choose of selection criteria of best models;
- evaluation of best model adequacy;
- use the final model.

Let us use the set theory to formalize the representation of data at each of these stages. We have the following components (built on the basis of analysis of structural identification [8]):

$W = (X, Y)$ - set of statistical data (sequence N values of random variable Y , which is characterized by M attributes X) $W = \{w_j\}, j = \overline{1, J}, J = n \cdot m, n = \overline{1, N}, m = \overline{1, M}$;

NW - set of normalized data, $NW = \{\bar{w}_j\}, j = \overline{1, J}$;

F - set of classes of models, $F = \{f_k\}, k = \overline{1, K}$;

G - set of generators of model structures, $G = \{g_l\}, l = \overline{1, L}$;

P - set of methods of parameter estimation, $P = \{p_r\}, r = \overline{1, R}$;

CR - set of criteria of model selection, $CR = \{cr_q\}, q = \overline{1, Q}$;

V - set of predicting models, $V = \{v_t\}, t = \overline{1, T}$.

Then the process of constructing the set of all possible models can be formally represented as a direct product of the sets components: $Z = W \times NW \times F \times G \times P \times CR \times V$.

Some element of the set Z described as

$$z_i = \{w_j, \bar{w}_j, f_k, g_l, p_r, cr_q, v_t\}, j = \overline{1, J}, k = \overline{1, K}, l = \overline{1, L}, r = \overline{1, R}, q = \overline{1, Q}, t = \overline{1, T},$$

$$i = \overline{1, I}, I = J \cdot K \cdot L \cdot R \cdot Q \cdot T,$$

is considered as specific data that was saved in the environment during the full cycle of building a particular model for given statistical data.

Thus, based on the apparatus of the set theory, a primary formalization was made for data saved in an XML-storage.

One can consider in details what information is stored after each of the system modules. At first, an import module places input statistical data and normalized data in the storage. The next stages of the process of models constructing from available statistical data, give information on the used model class, generator of model structures, method of parameters estimation and criteria for model selection, as well as on the built model, information about the structure and other characteristics of which is returned to the storage. Below we consider existing formats for saving predicting models and their use in our case.

4 Using PMML for storing information

Predictive Model Markup Language (PMML) is an XML-dialect used to describe statistical models and models of data mining. Its main advantage is that PMML-compliant applications can easily exchange models with other PMML-tools. The following classes of models can be kept using this markup language: associative rules, decision trees, center-based and distribution-based clustering, regression, general regression, neural networks, Bayes nets, sequences, text models, time series, rulesets, trees, support vectors.

In our case, for storing predicting models a scheme can be used which describes a regression function. Regression functions are used to determine the relationship between the dependent variable (target area) and one or more independent variables. The term regression usually refers to the prediction of numeric values, hence the PMML element

RegressionModel can also be used for classification. This is due to the fact that multiple regression equations can be combined in order to predict categorical values [9].

Using PMML allows to keep a simple regression model as follows:

$$\text{Dependent variable} = \text{intercept} + \text{Sum}_i (\text{coefficient}_i * \text{independent variable}_i) + \text{error}$$

Classification models can have multiple regression equations, in the next form

$$y_j = \text{intercept}_j + \text{Sum}_i (\text{coefficient}_{ji} * \text{independent variable}_i).$$

It should be noted that the authors use the most common version: PMML 3.2.

However, the use of PMML does not make it possible to store the full data set that should be kept in the storage. Therefore, in order to standardize the format preserving the full information it is better in this case to develop own XML scheme for storing all necessary information on the calculation results in a single format in this storage.

5 Using XML for storing information

Using XML to save data in storage provides sufficient opportunities for expansion of the system, including the possibility of developing proper schemes of XML-documents.

In the XML-storage input data should be saved, namely set of values of random variables and set of dependent parameters. Also, normalized data should be saved. It should be noted the ability to store not only predicting models but also an auxiliary information. Namely: information about the possible classes of models, generators of model structures, methods of parameters estimation, criteria of models selection.

Development of XML-schemes for storing information in the storage allows also generate reports not only for predicting models but for all the information contained in the repository.

6 Conclusion

This paper gives more detailed description of modules of integrated environment of information storing in the tasks of inductive modeling, including primary formalization of the information saving structures in XML-storage (input data and results of calculations).

Overview of the PMML use for storing results of calculations in the tasks of the GMDH-based inductive modeling was made. It is analyzed also the development of a proper XML-scheme for storing all necessary data used in forecasting tasks based on GMDH algorithms.

References

- [1] Ivakhnenko A. G., Stepashko V.S.: Noise-immunity of modeling. – Kiev: Naukova Dumka, 216 p, 1985.
- [2] Ivakhnenko A. G.: Inductive method of self organization of models of complex systems. – Kiev: Naukova Dumka, 216 p, 1982.
- [3] Shcherbakova N.V., Stepashko V.S. An Integrated Environment of Handling and Storing Information in the Tasks of Inductive Modeling // ICIM2008. – pp. 231-235.
- [4] Christopher J. Data: Introduction to databases systems. — K.: BHV, 608 p, 1998.
- [5] Date C.J.: An Introduction to Database Systems, Eighth Edition. – USA: Addison-Wesley, 1024 p, 2004.
- [6] Thomas K., Karely B.: Databases. Design, realization and accompaniment. Theory and practice. – M.: William, 1440 p, 2003.
- [7] Graves M.: Designing XML Databases. M: «Vil'yams» Publishing house, 640 p, 2002.
- [8] Stepashko V.S., Yefimenko S.M. Simulation experiments as a tool to study the effectiveness of modeling techniques for the observation data // USIM – 2008.
- [9] <http://www.dmg.org/> - Data Mining Group