INTEGRATED ENVIRONMENT FOR STORING AND HANDLING INFORMATION IN TASKS OF INDUCTIVE MODELLING FOR BUSINESS INTELLIGENCE SYSTEMS

Nataliya Shcherbakova, Volodymyr Stepashko

***Abstract**: Inductive modelling tools are widely used for solving problems of analysing economical, ecological, and other processes. Development of business intelligence systems based on inductive modelling algorithms for analysis, modelling, forecasting, classification, and clustering of complex processes is very promising.*

When solving real tasks of model construction from statistical data, the question of storage of and providing effective access to the information arises. At the stage of input data processing there are typical difficulties with processing data in different formats as well as containing omissions and untypically small or big values etc. From the other side, the question of output information storage exists like determination of structure and parameters of models, estimation of precision and validity, plots and diagrams drawing etc. This would allow structuring input data of different types and using the information already existing in database and also provide the storage of complete information on experiments and results of calculations.

To solve such kind of problems, the integrated environment for storing and handling information is developed. Architecture of the environment is offered giving the possibilities to manipulate present information freely using relational database containing only metadata and storing input statistical data and output results of calculations.

***Keywords**: integrated environment, handling and storing information, inductive modeling, GMDH-algorithms, Business Intelligence*

***ACM Classification Keywords**: H.2.8 Data Base Application – Data Mining*

Introduction

The number of companies which use business intelligence systems in their work is growing every year, so the development of such systems is continued; it is increasing constantly the number of systems, their functionality and technologies used for data processing and direct data analysis. Business intelligence systems are usually focused on

a specific task or function of a firm, such as analysis and forecast of sales, financial services, forecasting and risk analysis, trend analysis etc. Characteristic features of up-to-date BI systems are modularity, distributed architecture, the most common support and maintaining standards in web. At present, the development of systems aimed at business data analysis using OLAP, data mining and other. Growing range of algorithms is based on machine learning algorithms. Autonomous data mining tools are often included in other business analysis tools such as expanding database.

Analysis of algorithms used in business intelligence solutions

Business intelligence [Businessdictionary] refers to computer-based techniques used in spotting, digging-out, and analysing business data, such as sales revenue by products or departments or associated costs and incomes. But broader business intelligence can be characterized: firstly as process of converting data into information and knowledge for business decision support, secondly as information technology for data saving, information consolidating and guaranteeing access business users to knowledge, thirdly knowledge business gained as a result of data analysis and consolidation of information [Power, 2008]. Business intelligence solutions are widely used and have very diverse architecture, so there is no single specification of what those systems should be.

Objectives of a business intelligence exercise include understanding of a firm's internal and external strengths and weaknesses, understanding the relationship between different data for better decision making, detection of opportunities for innovation, and cost reduction and optimal deployment of resources.

For the past 10 years, the content and name of information-analytical systems have changed from information systems for manager, to decision support systems and business intelligence systems, second and third generation now. Business intelligence 2.0 is a new tools and software for business intelligence, beginning in the middle of 2000s, which enable, among other things, dynamic querying of real-time, web-based approached to data, as opposed to the proprietary querying tools that had characterized previous business intelligence software [Wikipedia]. Business Intelligence 3.0 is a term that refers to new tools and software for business intelligence, which enable contextual discovery and more collaborative decision making [Wikipedia].

Business intelligence tools is typically divided into the following categories: spreadsheets, reporting and querying software, OLAP, digital dashboards, decision engineering, process mining, business performance management, local information systems [Wikipedia]. Consider each of these groups in more detail.

Data mining is the process of extracting patterns from data; it is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. DM can be used to uncover patterns in data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. DM cannot discover patterns that may be present in a larger data body if those patterns are not present in a sample being "mined". Inability to find patterns may become a cause for some disputes between customers and service providers. Therefore data mining is not a proof fool but may be useful if sufficiently representative data samples are collected. The discovery of a particular pattern in a particular data set does not necessarily mean that a pattern is found elsewhere in the larger data from which that sample was drawn. An important part of the process is the verification and validation of patterns on other data samples.

Process mining is a process management technique that allow for the analysis of business processes based on event logs. The basic idea is to extract knowledge from event logs recorded by an information system. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs. Process mining techniques are often used when no formal description of the process can be obtained by other means, or when the quality of an existing documentation is questionable. For example, the audit trails of a workflow management system, the transaction logs of an enterprise resource planning system, and the electronic patient records in a hospital can be used to discover models describing processes, organizations, and products. Moreover, such event logs can also be used to compare event logs with some a priori model to see whether the observed reality conforms to some prescriptive or descriptive model.

Business performance management is a set of management and analytic processes that enable the performance of an organization to be managed with a view to achieving one or more pre-selected goals.

Local information systems (LIS) are a niche form of information system - indeed they can be categorized as Business intelligence tools designed primarily to support geographic reporting. They also overlap with some capabilities of Geographic Information Systems although their primary function is the reporting of statistical data rather than the analysis of geospatial data. LIS also tend to offer some common Knowledge Management type functionality for storage and retrieval of unstructured data such as documents. They deliver functionality to load, store, analyse and present statistical data that has a strong geographic reference.

Table 1. Most popular data mining algorithms offered by three different BI solutions

	Pentaho [Pentaho]	Microsoft BI [Microsoft]	Oracle BI [Oracle]
Decision Tree (DT)	Classifiers	Predicting a discrete or continuous attribute, Finding groups of common items in transactions	Classification
Linear Regression (LR)	Classifiers	Predicting a continuous attribute	
Naive Bayes (NB)	Classifiers	Predicting a discrete attribute	Classification
Clustering	Clusterers	Predicting a discrete attribute, Finding groups of similar items	
Association Rules (AR)	Classifiers	Finding groups of common items in transactions	
Sequence Clustering (SC)	Forecasting	Predicting a sequence	
Time Series (TS)	Forecasting	Predicting a continuous attribute	
Neural Network (NN)	Forecasting	Predicting a continuous attribute	
Support Vector machine (SVM)	Classifiers		Classification and Regression
One Class Support Vector Machine (One-Class SVM)			Anomaly Detection
Generalized Linear Models (GLM)			Classification and Regression
Minimum Description Length (MDL)			Attribute Importance
Apriori (AP)	Association		Association
k-Means (KM)	Clusterers		Clustering
Orthogonal Partitioning Clustering (O-Cluster or OC)			Clustering
Nonnegative Matrix Factorization (NMF)			Feature Extraction

Business intelligence systems implement often classical data mining algorithms [Thomas, 2003]. Classification algorithms predict one or more discrete variables based on the other attributes in the dataset. Regression algorithms predict one or more continuous variables such as profit or loss on the bases of other attributes in the dataset. Segmentation algorithms divide data into groups or clusters of items that have similar properties. Association algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating

association rules can be used in a market basket analysis. Sequence analysis algorithms summarize frequent sequences or episodes in data such as a Web path flow.

Furthermore, most business intelligence systems implement provisional data algorithms allowing eliminate gaps, specifically small or large data and convert output data to a required format.

Classical algorithms are mostly implemented as standard library for use in the systems being developed. We have examined main characteristics of the most popular business intelligence systems, namely the three different software packages: Pentaho, Microsoft BI, and Oracle BI. Table 1 shows the most known data mining algorithms offered by these three different BI solutions. It should be noted that Weka (Pentaho Data Mining) has near 100 classification schemes.

Prospects of inductive modelling algorithms usage in Business Intelligence solutions

Algorithms of inductive modelling are applicable for solving real-world modelling tasks in economical, ecological, and other processes [Ivakhnenko, 1985], [Ivakhnenko, 1982]. They are widely used in ill-defined systems developed for solving a specific problem. Below we examine a question of possibility of their use in business analytics; namely, which algorithms of inductive modelling and for what purposes can be used in business intelligence solutions.

The article [Ivakhnenko, 1968] published in 1968 by Prof. O.G. Ivakhnenko has marked the beginning of the new scientific direction called "inductive self-organizing of models from experimental data" or simply "inductive modelling" [Stepashko, 2008].

Group method of data handling (GMDH) is a family of inductive algorithms for computer-based mathematical modelling of multi-parametric datasets that features fully-automatic structural and parametric optimization of models. GMDH as personification of the inductive approach is an original method for constructing models from experimental data under uncertainty conditions. Models of optimal complexity obtained by this method reflect unknown laws of functioning of an object (process) information about which is implicitly contained in a data sample. To build models automatically, GMDH applies the principles of variants generation, nonterminal decisions and successive selection of the best models according to external criteria. These criteria are based on dividing the data set into two parts, where the tasks of parameter estimation and model checking are implemented in various subsets. GMDH algorithms are characterized by an inductive procedure that performs sorting-out of gradually complicated polynomial models and selecting the best solution using the external criteria.

A GMDH model with multiple inputs and one output is a subset of components of the base function:

$$Y = Y(x_1,\ldots,x_m) = a_0 + \sum_{i=1}^{m} a_i f_i(x_1,\ldots,x_m) \ , \qquad (1)$$

where *f* are elementary functions of different sets of inputs, *a* are coefficients and *m* is the number of the base function components.

To find the best solution, GMDH algorithms consider various component subsets of the base function (1) called partial models. Coefficients of these models are estimated by the least squares method. GMDH algorithm gradually increases the partial model complexity and finds an optimal model structure and parameters indicated by the minimum value of an external criterion. This process is called self-organization of models.

The most common base function used in GMDH is the Kolmogorov-Gabor polynomial:

$$Y(x_1,\ldots,x_m) = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=i}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=i}^{m}\sum_{k=j}^{m} a_{ijk} x_i x_j x_k + \ldots \qquad (2)$$

GMDH is used in such fields as data mining, knowledge discovery, prediction, complex systems modelling, optimization and pattern recognition. Application of GMDH algorithms for solving forecasting tasks allows their use in business intelligence systems along with other data mining algorithms.

Among GMDH algorithms that have more widespread gained, we can indicate the following [Ivachnenko, 2007]: Combinatorial (COMBI), Multilayered Iterative (MIA), GN, Objective System Analysis (OSA), Harmonical, Two-level (ARIMAD), Multiplicative-Additive (MAA), Objective Computer Clusterization (OCC), “Pointing Finger” (PF) Clusterization Algorithm, Analogues Complexing (AC), Harmonical Rediscretization, Algorithm on the base of multi-layered Theory of Statistical Decisions (MTSD), Group of Adaptive Models Evolution (GAME).

It should be noted that algorithms proposed by the most popular BI solutions are used for tasks of classification and clustering, mostly. For the numerical forecasting, developers offer: neural networks, linear regression and model trees. Such tools are less winning in forecasting tasks than GMDH algorithms. Requests for functionality of business intelligence system today lie in expanding their capacity of data mining. Other subsystems of business intelligence, such as data integration, import-export and reporting tools are developed very quickly. Thus the direction of business intelligence systems development using GMDH-based forecasting algorithms is very promising. Attention should also be drawn to the development of inductive modelling algorithms for classification and

clustering. Effective using such algorithms for modelling of complex systems is also an important argument for applying them in business intelligence systems.

Given the above we are developing our own system in which we use GMDH algorithms for modelling complex systems (processes). A necessity in accessible storage and drawing on scientific researches ripened already a long ago. An integrated environment for information storage would help to solve the existing problems, allowing structuring input data of different types and sources already existing in a data base, and also providing storage of complete information on experiments and results of calculations.

Integrated environment for storing and handling information

In [Shcherbakova, 2008] an architecture of integrated environment of handling and storing information in the tasks of inductive modelling was proposed, which allows freely manipulate the available information through building a layout environment which consists of relational database [Christopher, 1998], [Date, 2004], [Thomas, 2003] containing only meta-data and XML (PMML) storage, in which input data and results of calculations are stored [Graves, 2002].

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

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.

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 an XML storage. Auxiliary information such as data on the user, date and time use of files etc. it is better to store in a relational database.

Below we consider existing formats for saving predicting models and their use in our case.

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.

for storage of predicting models, a scheme can be used in our case 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.

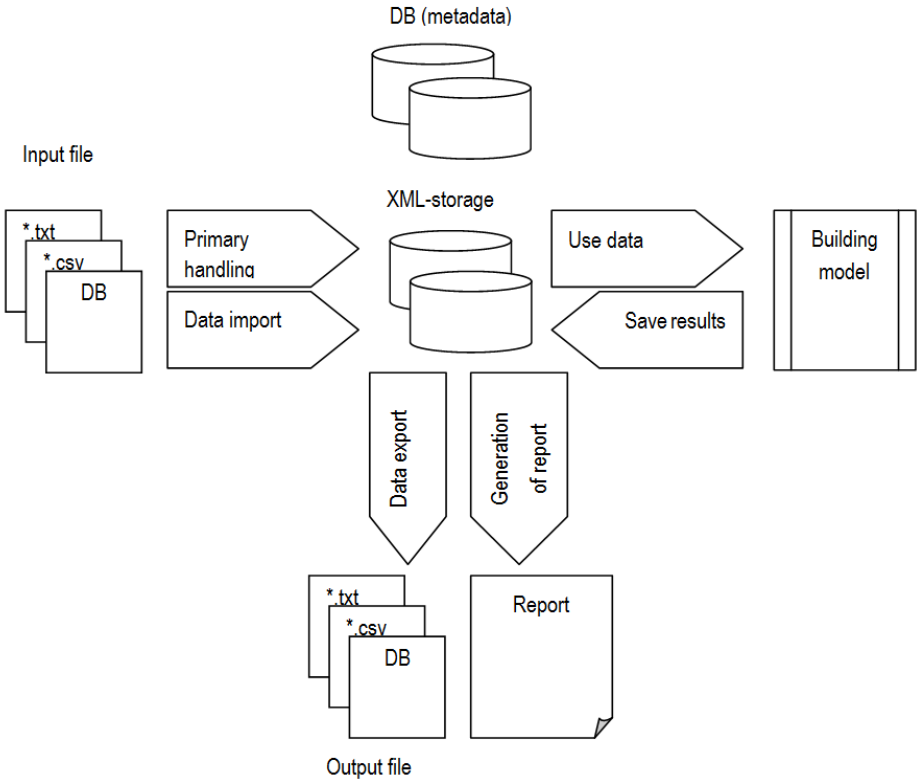

Fig. 1 Architecture of an integrated environment for information storing and handling

The offered integrated environment is intended for working out problems of storage of input statistical data and handling results. Developed architecture of the system of information storage in the tasks of inductive modelling gives a possibility to develop a software system that will allow freely manipulating the available information and adding new one.

Conclusion

This paper presents project of an integrated environment for storing and handling information in tasks of inductive modelling based on algorithms of the Group Method of Data Handling. Such type of system can be applied to solve some tasks of business intelligence, including forecasting, classification and clustering. Applying GMDH algorithms in business intelligence systems gives a particularly promising opportunity towards building complex models for business data analysis.

Bibliography

[Businessdictionary] http://www.businessdictionary.com/

[Power, 2008] Power D.J.: A Brief History of Decision Support Systems, version 4.0. DSSResources.COM. Retrieved, 2008.

[Wikipedia] http://en.wikipedia.org/wiki/Business_Intelligence_2.0/

[Wikipedia] http://en.wikipedia.org/wiki/Business_intelligence_3.0/

[Wikipedia] http://en.wikipedia.org/

[Thomas, 2003] Thomas K., Karely B.: Databases. Design, realization and accompaniment. Theory and practice. – M.: William, 1440 p, 2003.

[Pentaho] http://wiki.pentaho.com/ - Pentaho documentation.

[Oracle] http://msdn.microsoft.com/ - MSDN.

[Microsoft] http://docs.oracle.com/ - Oracle documentation.

[Ivakhnenko, 1985] Ivakhnenko A.G., Stepashko V.S.: Noise-immunity of modelling. – Kiev: Naukova Dumka, 216 p, 1985.

[Ivakhnenko, 1982] Ivakhnenko A.G.: Inductive method of self organization of models of complex systems. – Kiev: Naukova Dumka, 216 p, 1982.

[Ivakhnenko, 1968] Ivakhnenko O.G.: Group method of data handling - rival of method of stochastic approximation. //Automatic №3. – Kiev, pp 58-72, 1968.

[Stepashko, 2008] Stepashko V.S.: Theoretical aspects of GMDH as a method of inductive modelling. – Proceedings of the II International Conference on Inductive Modelling ICIM-2008, 15-19 September 2008, Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, pp 9-16, 2008.

[Shcherbakova, 2008] Shcherbakova N. and Stepashko V. Integrated Environment for Information Handling and Storage in the Tasks of Inductive Modeling. – Proceedings of the II International Conference on Inductive Modelling ICIM-2008, 15-19 September 2008, Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, pp 231-235, 2008.

[Ivakhnenko, 2007] http://www.gmdh.net/

[Christopher, 1998] Christopher J. Data: Introduction to databases systems. — K.: BHV, 608 p, 1998.

[Date, 2004] Date C.J.: An Introduction to Database Systems, Eighth Edition. – USA: Addison-Wesley, 1024 p, 2004.

[Thomas, 2003] Thomas K., Karely B.: Databases. Design, realization and accompaniment. Theory and practice. – M.: William, 1440 p, 2003.

[Graves, 2002] Graves M.: Designing XML Databases. M: «Vil'yams» Publishing house, 640 p, 2002.

http://en.wikipedia.org/wiki/Predictive_Model_Markup_Language/

Authors' Information

Nataliya Shchrebakova *– PhD student of IRTC ITS of NASU, P.A.: 40, Akademik Glushkov Prospect, Kyiv, Ukraine, 03680; e-mail: nataliya.shchrebakova@gmail.com*

Main Fields of Scientific Research: Information technologies of inductive modelling, Business Intelligence solutions

Volodymyr Stepashko *– Head of Department for Information Technologies of Inductive Modeling of IRTC ITS, Professor, Dr Sci, P.A.: 40, Akademik Glushkov Prospect, Kyiv, Ukraine, 03680; e-mail: stepashko@irtc.org.ua*

Main Fields of Scientific Research: Data analysis methods and systems, Knowledge discovery, Information technologies of inductive modelling, Group method of data handling (GMDH)