Data Mining Methods Application to the Problem of Handling Corporative Dataset on Heavy Oil Production

Iakov S. Korovin, Maxim V. Khisamutdinov, Anatoly I. Kalyaev
Scientific Research Institute of Multiprocessor Computer Systems of the Southern Federal University
347928, Russian Federation
korovin_yakov@mail.ru

***Abstract*—In this paper, we perform the analysis of Data Mining methods; the application of those is decided to be the most effective in the task of handling information on the production processes, typical to the fields of heavy oil in Western Siberia. The common feature of heavy oil production in Russia is low efficiency of classical math models applied in the industrial software, aimed to reduce the oil production cost. Thus, it demands the industrial usage of novel approaches, where among the most probable ones are the artificial intelligence technologies. The summary of their classification with the advantages' analysis is presented.**

Keywords—Data Mining, heavy oil production, neural networks, decision trees, genetic algorithms, fuzzy logic

I. DATA MINING METHODS VARIETY AND THEIR APPLICABILITY IN THE TASK OF HEAVY OILPRODUCTION DATABASE INFORMATION HANDLING

Modern digital oilfield framework implements the application of Data Mining techniques. Data Mining is being understood as "production" or "data retrieval". Quite often near Data Mining the words "detection of knowledge in databases" (Knowledge Discovery in Databases) and "the intellectual analysis of data" are met. They can be considered as synonyms of Data mining. Emergence of all specified terms is connected with a new round in development of means and methods of data processing [1].

There is a set of definitions of Data mining, but in general, they coincide in allocation of four main signs: Data Mining is a process of detection in crude data

- earlier unknown,
- uncommon,
- practically useful,
- knowledge (regularities) available to interpretation, necessary for decision-making in various spheres of human activity.

Results of Data mining are empirical models, classification rules, the allocated clusters, etc. - it is possible to incorporate them in the existing decision support systems and to use them for the future situations forecast.

Finding of the hidden rules in data, interrelations between various variables in databases, modeling and studying of difficult systems on the basis of their behavior history –these are the subject and tasks of Data mining science [1].

The basis of the Data mining technology is made by the concept of the templates, representing regularities [1]. To various regularities types there correspond certain tasks of Data mining:

- classification,
- clustering,
- forecasting,
- association,
- visualization,
- analysis and detection of deviations,
- estimation,
- analysis of communications,
- summing up.

Clustering

Feature of a clustering is that classes of objects are initially not predetermined. Splitting objects into groups is result the given procedure.

Forecasting

As a result of the forecasting on the basis of historical data features the future values of target numerical indicators are estimated. Methods of mathematical statistics, neural networks, etc. are widely applied to the solution of such tasks.

Association

While solving a problem of associative rules search, regularities between the connected events in a dataset are found. Difference of association from two previous tasks is that search of regularities is carried out not on the analyzed object basis and between several events which occur at the same time. The most known algorithm of associative rules search - algorithm of apriority. The sequence allows to find temporary regularities between transactions. The problem of sequence is similar to association, but its purpose is

establishment of regularities not between at the same time coming events, and between the events, connected in time (i.e. happening to some certain time interval). In other words, the sequence is defined by high probability of a chain of the events connected in time. Actually, the association is a special case of sequence with "a temporary log" (time lag).

Visualization

As a result of visualization the graphic image of the analyzed data is created. For the solution of a problem of visualization the graphic methods, showing existence of regularities in data are used. Example of methods of visualization - data presentation in 2-D and 3-D measurements.

Analysis and detection of deviations

Applied for detection and analysis of the data, most different from the general dataset, identification of so-called uncharacteristic templates.

Estimation

The task is reduced to a prediction of continuous values of a sign.

Analysis of communications

It is a problem of finding of dependences in a dataset.

Summing up

It is a task, which purpose is the description of concrete groups of analyzed data set objects.

Table 1 Data mining methods

№	Method	Short characteristics	Estimation of method, disadvantages
1	**Statistical methods**	1. Classical methods of the multidimensional statistical analysis (correlation and regression, factorial, etc.) allow to solve the widest class of statistical tasks; there is a wide range of software. 2. Nonlinear regression methods are perspective, give opportunity of the group accounting of arguments, provide statistically significant results	Despite comparative efficiency, are impossible for application without knowledge in the domain and incompleteness or absence of statistical data.[2-5]
2	**Fuzzy logic methods**	It is successfully applied in the solution of tasks, in which basic data is fuzzy (inexact, incomplete, contradictory, distorted, noisy). Advantages: the description of conditions and a method in the language close to natural, universality (according to the theorem – any mathematical system can be approximated by the system founded on fuzzy logic). The solved tasks: classification and the data analysis, inference in the conditions of uncertainty and problems of decision-making	The initial set of the postulated fuzzy rules is formulated by the expert, this set of rules can be incomplete or contradictory; the view and parameters of the membership functions, describing input and output variables are formed subjectively and can badly reflect reality
3	**Genetic algorithms**	Methods represent operations, modeling evolutionary processes on the basis of genetic inheritance and selection mechanisms[6]	Complexity of initial model creation
4	**Neural network approach**	Non-formalizable or fuzzy tasks with modeling of difficult nonlinear dependences between factors and target indicators, identification of tendencies in the input generalizing dependences, receiving substantial results – are solved at rather small volume of initial information, the subsequent specification of models (retraining) is possible. Allow to solve problems of clustering, classification and image recognition, functions approximation, prediction/forecast, optimization [2-5]	Doesn't demand big computing resources. Possesses a high speed of data processing. Shortcoming: complexity of an explanation of the made decision.
5	**Fuzzy neural networks**	Networks are adaptive, allow adjustment in the course of work. Formally on structure are identical to a multilayered neural network with training	Practically don't remove shortcomings of methods of fuzzy logic. Are difficult in formation of the training sets
6	**Fuzzy situational inference**	The limited set of fuzzy situations can describe almost infinite number of conditions of a controlled object. Advantages: flexibility and resistance to unforeseen changes of development and decision-making conditions, sufficient simplicity of algorithms of an fuzzy situational inference	Identified input fuzzy situation is compared to all standard situations. Demands huge computing resources. Big complexity in initial model creation.
7	**Cognitive maps**	Cognitive mapping are the graph mathematical models intended for formalization of the description of difficult object, a problem or system functioning and identification of relationships of cause and effect between their elements as a result of impact on these elements or changes of communications nature. The main advantage – possibility of application methods of cognitive modeling in other methods at different stages.	Complexity of definition on the aim set, optimum strategy its achievement; lack temporary parameters modeling possibility.
8	**Decision trees**	Method is suitable only for the solution of problems of classification and partially-for the solution of numerical forecast tasks. Advantage: presentation of rules and, transparent inference.	The main shortcoming – realization of this method demands huge computing resources.
9	**Hierarchical methods of the analysis and decision-making**	Methods are intended for finding of alternative decisions on the basis of synthesis of multiple judgments and receiving system of preferences. Are based on hierarchical representation of the elements of system, defining an essence of any problem. Relative extent of interaction of elements in hierarchy is defined.	Severe dependence on the expert and the decision-maker

II. CONCLUSIONS

In whole, the situation within Russian oil and gas industry is rather difficult. The importance of the heavy oil production cost reduction becomes more and vital in the conditions of “ordinary” oil fields stores becoming quite low day after day and taking into consideration the current ‘era of the cheap hydrocarbons’. One of the most possible approaches is to apply new methods of data processing in the procedures of the heavy oil enhancement of recovery. The research, aimed on the analysis of the modern Data Mining techniques, is conducted by the authors. The solution is to apply fuzzy modeling of the heavy oil production processing, based on the handling of the historic dataset with the hybrid usage of artificial neural networks, evolutionary algorithms and other novel methods.

ACKNOWLEDGMENTS

This paper is published due the financial support of the Russian Science Foundation (RSF) via the grant № 15-19-00196.

III. REFERENCES

[1] A.A. Bargesyan, M.S. Kupriyanov and others. /Data analysis technologies. Data Mining, Visual Mining, Text Mining, OLAP// Published: BHV-Petersburg, 2007, 384 pp.;

[2] I.S. Korovin, M.V. Khisamutdinov. Neuronetwork decision support system for oilfield equipment condition online monitoring. Advanced Materials Research (Trans Tech Publications, Switzerland). Volume 902 (2014), Pages 409-415.

[3] I.S. Korovin, M.V. Khisamutdinov. Hybrid method of dynamograms wavelet analysis for oil-production equipment state identification. Advanced Materials Research (Trans Tech Publications, Switzerland. Volume 909 (2014).

[4] Korovin, Ya.S. Decision support system for electrical submersible pumps control on the neural network basis. Neftyanoe khozyaystvo - Oil Industry. Issue 1, January 2007, Pages 80-83.

[5] Korovin, Ya.S., Tkachenko, M.G., Kononov, S.V. Oilfield equipment's state diagnostics on the basis of data mining technologies. Neftyanoe khozyaystvo - Oil Industry. Issue 9, September 2012, Pages 116-118

[6] Korovin. I.S; Khisamutdinov M.V.; Kaliaev A.I. The Application of Evolutionary Algorithms in the Artificial Neural Network Training Process for the Oilfield Equipment Malfunctions' Forecasting. PROCEEDINGS OF THE 2ND INTERNATIONAL SYMPOSIUM ON COMPUTER, COMMUNICATION, CONTROL AND AUTOMATION Book Series: Advances in Intelligent Systems Research. Volume: 68. Pages: 253-257. Published: 2013