Cognitive methodology for forecasting oil and gas industry using pattern-based neural information technologies

O Gafurov1, D Gafurov2, V Syryamkin1
1National Research Tomsk State University, Tomsk, Russia
2Krasnoyarsk Research Institute of Geology and Raw Materials, Krasnoyarsk, Russia
E-mail: oleg.gafurov@mail.ru

Abstract. The paper analyses a field of computer science formed at the intersection of such areas of natural science as artificial intelligence, mathematical statistics, and database theory, which is referred to as "Data Mining" (discovery of knowledge in data). The theory of neural networks is applied along with classical methods of mathematical analysis and numerical simulation. The paper describes the technique protected by the patent of the Russian Federation for the invention “A Method for Determining Location of Production Wells during the Development of Hydrocarbon Fields” [1–3] and implemented using the geoinformation system NeuroInformGeo. There are no analogues in domestic and international practice. The paper gives an example of comparing the forecast of the oil reservoir quality made by the geophysicist interpreter using standard methods and the forecast of the oil reservoir quality made using this technology. The technical result achieved shows the increase of efficiency, effectiveness, and ecological compatibility of development of mineral deposits and discovery of a new oil deposit.

1 Introduction
The discovery-driven data mining technology presented in the article is based on the concept of patterns proposed the authors; this concept reflects many aspects of data relationships. Patterns represent regularities inherent in subsamples of data and they can be compactly expressed in a human-readable form.

In the NeuroInformGeo geoinformation system, the technique of the intellectual analysis of geological, geophysical and geochemical parameters based on the neural network methods with ranking of input attributes by the significance level is realized. The significance of the input characteristic provides an assessment of the informative contribution of this attribute to the result of the neural network training and realization of the forecast by the trained neural network.

The technology presented assumes the implementation of tuning and training of neural networks on the etalon field, which was brought into development with the appropriate geological complex and representative array of geological and geophysical data, which describe this etalon complex. The neural networks trained on the etalons form a pattern that can be used at the stages of exploration and additional exploration of oil and gas deposits in productive horizons of already discovered deposits. Taking into account the statistical data, they also can be used at the search stage on unexplored and poorly explored areas.
2 Pattern of the geological and geophysical complex

Accumulation of the experience in forecasting and interpretation can be ensured in the form of formation, preservation, and correct application of patterns that represent the library of trained neural networks on etalon deposits or drilled areas [4].

2.1 Pattern determination

A pattern is the accumulated experience of forecasting in the form of a library of trained neural networks on etalon objects.

Regarding to geological prospecting, reference objects are deposits or drilled areas that is objects with such a level of exploration, which allows checking the results of the neural network training using the reliable data on the geological structure of the etalon object.

2.2 Formation of the pattern of the geological and geophysical complex

The study of the entire volumetric part of the geo-environment is performed by geophysical methods (seismic and magnetic prospecting, etc.). During the process of natural fields studying (magnetometry, gravimetry, etc.) or based on compulsorily excited fields (seismic prospecting and electrical exploration), a three-dimensional matrix is formed where real geological objects are replaced by digital images of their physical fields.

Let us introduce the concept of a physical model of the medium in the form of a function $Q_n(x, y, z)$, which describes the distribution of a physical parameter in $A$ region, which creates a geophysical field $f_n(x, y, z)$ in the surrounding space with geomedia parameters $p_1, \ldots, p_k$, and also the relation:

$$f_n(x, y, z) = f(Q_n, p_1, \ldots, p_k),$$ (1)

It connects $Q_n$ and $f_n$, that is allows calculating the field of the given geophysical parameter (the direct problem of geophysics).

To calculate the field of the geophysical parameter $f_n(x, y, z)$, we need the values of $Q_n(x, y, z)$ and parameters of the geo-environment $p_1, \ldots, p_k$, as well as the position of the boundary (contact) surfaces. In seismic prospecting, this includes the value of the rock density and the boundary of their differences; in geology, this includes the value of clay and sandstone density, i.e. the density value of geological bodies and objects.

The problem of constructing a geological model $Q_n(x, y, z)$ using the geophysical parameter $f_n(x, y, z)$ measured is called the inverse problem. It is necessary to find distribution of parameters of the geo-environment $p_1, \ldots, p_k$ and position of the contact surfaces.

Let us imagine that formation of answers for each step is a statement of $P(1), P(2), P(3), \ldots$, each of which can be either true or false, i.e. it may bring us to the problem solution or lead away. The principle of finite induction states that in order to prove the truth of propositions for all $P(n)$ it is sufficient to establish the truth of $P(1)$ and the truth of an infinite sequence of implications:

$$P(1) \Rightarrow P(2) \Rightarrow P(3) \Rightarrow \ldots P(n) \Rightarrow P(n+1) \Rightarrow \ldots$$ (2)

The machine must correctly calculate the parameters at every step; and the initial problem must be set appropriately to a real geological environment. On each cycle, a neurocomputer program calculates the parameters based on the training sample.

2.3 An example of creating a pattern based on the etalon deposit in East Siberia

At the first stage of forming trained neural networks (patterns), an etalon deposit is determined, which was brought into development (Figure 1). Therefore, it has a representative array of geological and geophysical information, which will allow correct training and testing of neural networks and creating a pattern for further forecast of field and geological parameters in nearby promising areas (licensed areas - LA) poorly explored by exploratory drilling.

When forming a pattern in NeuroInformGeo geoinformation system [5], all input parameters (maps (meshes) of seismic attributes, geophysical, geochemical and other parameters) are normalized to the range [-1...1]. Before feeding to the synapse, each input signal is recalculated according to the equation:
\[ Y_i = 2 \times (x_i - \text{min}_i) / (\text{max}_i - \text{min}_i) - 1 \]  \hspace{1cm} (3)

where \( x \) is the initial signal, \( Y_i \) is the received normalized signal, \( \text{min} \) and \( \text{max} \) are respectively the minimum and maximum values of the input parameters interval in the field fed to the synapse \( i \).

During the neural network training, the data studies of wells are used, in which the tests and a complex of geophysical studies (GIS) were performed. A multilayer neural network is trained according to the backpropagation algorithm (backward propagation of errors). The network architecture is selected empirically providing a minimum error.

For each of the input parameters, the significance indicators are determined, which characterize the informative contribution of this parameter (attribute) to the results. When solving the \( q-o \) example, the significance indicator of the parameter is determined by the following formula:

\[ \chi_{N_a}^q = \left| \frac{\partial H}{\partial W_{N_a}} (w_{N_a} - w_{N_a}^*) \right| \]  \hspace{1cm} (4)

where, when solving the \( q-o \) example, the significance indicator shows how much the value of the estimation function changes in case of solving the \( q-o \) example by the network if the current value of the parameter \( W_{N_a} \) is replaced by the nearest selected value \( W_{N_a}^* \) for the \( N_a \) parameter. The final significance indicator of the \( N_a \) parameter is calculated as the total average:

\[ \chi_{N_a} = \frac{1}{n} \sum_{q=1}^{n} \chi_{N_a}^q \]  \hspace{1cm} (5)

where \( n \) is the number of examples. Thus, the calculated value of the significance indicator for the \( N_a \) parameter essentially represents (in the linear approximation) the absolute value of the change in the estimation function when the input parameter (signal) is removed from the network. Non-informative parameters are removed from the network, the neural network is retrained.

**Figure 1.** Zones of a possible forecast of the reservoir quality based on the Pattern created using seismic data and neuro-information technologies. The first zone means formation of etalon neural networks for productive horizons within the geological complex studied (Pattern). The second zone is a forecast of the reservoir quality (effective thickness, productivity, and permeability) on poorly explored licensed area. The third zone means the zone where it is possible to forecast the quality of the collectors based on the etalon pattern created.
The process of multiple iterations results in calculation of a true mean error of the neural network generalization.

If the training results are unsatisfactory, the initial geological and geophysical data are rechecked and the neural network is reset based on another feature that also characterizes the environment under the study.

If the solution to the problem is satisfactory, a pattern is formed, which is further used to forecast geological and field parameters on poorly explored LAs associated with an etalon deposit (based on which the pattern was created) by general geological and genetic characteristics.

Forecasting is carried out according to the etalon set of parameters corresponding to the pattern. The program classifies the multidimensional feature space into classes of membership (similarity) to a certain etalon and generates a map of etalons for the area interpretation or forms a cut with distinguishing complexes (classes) as temporary or deep models.

Figure 2 shows one of the patterns created. It characterizes the established dependence of the values of the effective thicknesses of Osinsky productive horizon (Nepa arch) on geophysical parameters (a complex of attributes of the seismic wave field), which describe the geological environment very accurately thus allowing training the neural network to forecast characteristics of the productive layer quality.

![Figure 2. The pattern characterizing the established dependence of the values of the effective thicknesses of Osinsky productive horizon (Nepa arch) on the complex of attributes of the seismic wave field.](image)

Training of the neural network was conducted in the borehole environment of exploratory wells according to a set of seismic attributes. In addition to the effective thicknesses ($N_{eff}$) of the reservoir, the predicted parameters also included the porosity coefficient ($K_{por}$), the permeability coefficient $\nu$, and various inflow characteristics. One of the most obvious classifications for inflows is carried out based on hydrodynamic studies conducted in exploratory wells (HDSW) that characterize the inflow type (lack of inflow, water, gas, oil). However, this classification, like the flow rate classifier, largely depends on the quality of drilling and tests conducted in different periods of the deposit exploration. For each predicted parameter, a neural network with layered architecture was created; it included up to 8–10 layers of 50–100 neurons in each hidden layer.
The training was conducted using the conjugate gradient method. Training using the conjugate gradient method showed the fastest convergence in estimating minimization of the error function (with a given accuracy).

During the training to predict the $N_{ef}$ parameter, the neural network learned to recognize (predict) the values in the wells in 985 training cycles; besides, with a given accuracy of 1 meter, the average error was 0.6 meters, the maximum error was 0.9 meters, and all examples were solved correctly. When configuring a neural network to recognize $K_{por}$ in the well area, all the examples included in the training sample were also solved correctly with a given accuracy of 1%. During the classification according to the HDSW data, there were no contradictory examples in the training sample and all examples were solved.

Testing of the trained neural network by cross-validation on test wells makes a big error. For example, when predicting $N_{ef}$, the average forecast error is 2.9 m, and the maximum error is 6.4 m (Table 1).

| Well number | $N_{ef}$ according to GIS (m) | Average value of the forecast $N_{ef}$ (m) $R=50$ m |
|-------------|-------------------------------|---------------------------------|
| 67          | 0                             | 2.7                             |
| 27          | 5.2                           | 6.7                             |
| 51          | 6                             | 6.3                             |
| 103         | 7.2                           | 7.6                             |
| 59          | 8.8                           | 12.4                            |
| 72          | 9.2                           | 11.4                            |
| 75          | 11.4                          | 11.2                            |
| 122         | 20.8                          | 15.5                            |
| 40          | 4                             | 3.8                             |

Average error: 2.9
Max error: 6.4
Max scatter: 2.8

However, this range of error values is acceptable, taking into account that the forecasting is performed using a complex of seismic attributes, the vertical calculation range of which reaches 0.03 s. and more, which is because of the oscillation period length. Accordingly, the resolving power of the forecast depends on the wave properties of the time section and the method of measuring seismic characteristics and is determined primarily by the seismic parameter analysis range. At the same time, for some attributes, the calculating range exceeds the capacity of the Bilirsky Series and hence the productive horizon, which is of interest to us, because it covers thickness of not only a productive horizon, but also overlying impermeable layers of rock salt.

Thus, in a complex interpretation, the informative contribution of these attributes is based on an integral approach. The physical meaning of the integral analysis of dynamic attributes of the seismic wave field is presence of the seismic field in section of anomalous objects. The models of environments characterized by local unconsolidated areas, stress state of rocks, hydrocarbon dispersion, tectonic disturbances, and reef formations meet the requirements for forming anomalies [6,7].

Figure 3 shows the dependence of the predicted $N_{ef}$ values of the effective thicknesses of the reservoir on the $N_{ef}$ values in the wells according to GIS (numbers of test wells are signed). The dependence is linear, the square of the linear regression coefficient is 0.78, taking into account all the values predicted by the neural network in a certain confidence radius around the well. If considering only the average predicted $N_{ef}$ values, the value of the square of the linear regression coefficient increases and amounts to 0.84 (correlation coefficient 0.92) on the average (for different sets of training and testing samples), which is excessive accuracy for solving seismic data prediction
problems and sufficient for the pattern created. In this case, the maximum error of 6.4 m appears in the wells with the maximum effective thickness values determined according to the GIS data.

Thus, the neural network slightly understates the maximum values of the effective thickness in the wells with the maximum $N_{ef}$ values. However, in general, it correctly divides the wells in quality predicting zones with no reservoirs as well as predicting the $N_{ef}$ values (in meters) in the test wells and other parameters characterizing the reservoir quality.

**Figure 3.** Dependence of the forecast $N_{ef}$ values on the $N_{ef}$ values in the wells according to the GIS data.

In order to understand the physical nature of the possibility of creating a pattern and forecasting the quality of reservoirs, we determined the attributes making the most significant contribution to the training and forecast of each parameter.

In general, when analyzing the forecast parameters, coefficients of the attribute significance were determined (as a result of the neural networks training in different productive horizons). Energy of scattered waves, acoustic impedance, and frequency and amplitude characteristics of the seismic wave field make a significant contribution to the forecast of effective thicknesses.

Scattered waves occur at supercritical angles of reflection from diffraeters located in rock mass. Fractures, cracks, and caverns can be such diffraeters. Acoustic impedance is a product of velocity in a layer and density. Thus, the change in impedance is associated either with a change in the propagation velocity of longitudinal waves, or with a change in density, and this relationship is linear. Besides, the velocity decreases with increasing porosity, increases with increasing pressure, decreases with increasing temperature, and decreases with fluid saturation.

For perspective strata, during the process of determining the reservoir quality, the lateral variation of the dominant frequency $F_1$ and the amplitude of the value of the scattered component in the analysis range also play a significant role in the forecast. Lateral changes in the dominant frequencies $F_1$, $F_2$, $F_3$ may indicate a fluid saturation or increased fracturing, which may occur because of changes in stratigraphy, lithology, and fault tectonics. As an example, Figure 4 shows the position of the dominant frequencies on the amplitude-frequency spectrum. The attribute of instantaneous frequency and instantaneous amplitude is obtained because of the Hilbert transformation of the original seismic trace. However, the physical meaning of the attributes after the Hilbert transformation does not change.
For promising strata, the effective width of the amplitude-frequency spectrum also makes a significant contribution; the change of this parameter depends on the change in the dominant frequencies on the amplitude-frequency spectrum.

Amplitude parameters make a significant contribution to the forecast for all layers. The change in the amplitudes of the seismic traces corresponds to a change in the reflection coefficients. It follows from the Zoeppritz equation that under normal incidence of the P-wave on the boundary, no exchange waves appear, and the reflection coefficient is determined by the known impedance relation. However, in the case of a non-normal incidence, the Poisson's ratio plays an important role, or simply the ratio of the longitudinal velocity to the transverse velocity. In general, a change in the amplitude of reflected waves may indicate a change in lithology, density, fluid saturation, porosity, fracturing, and others.

The anomalies of the most significant attributes are confined to the layers of the positive structure in the context of the standard 3D cube, which may reflect the nature of the coastal-marine sedimentation conditions. The determined anomalies of the 3D cube coincide with the most significant attributes of the focusing transformation established during the training on the \( N_{ef} \) parameter - the maximum positive and mean square amplitudes of the scattered component of the seismic field. In the physical sense, this attribute reflects the most probable zones of decompression of the medium expressed in the increase of fracture and voidage. Attributes of the focusing transformation [8] also make a great contribution to the forecast of porosity and inflow classes.

In general, creation of the pattern resulted in forming the neural networks, which were trained to forecast the following parameters on the territory of the Nepa arch:
- effective thickness of layers;
- porosity coefficient;
- the nature of saturation (absence, water, oil, gas);
- effective pore volume;
- coefficient of permeability

The pattern was formed based on the data of exploratory wells (more than 100 wells) and 3D seismic data of the studied deposit of the Nepa arch in Irkutsk region (Russia).

2.4 Application of the Pattern on the licensed areas

On the licensed areas (LA) neighboring to the etalon deposit (Pattern) and located within the boundaries of Nepa arch in Irkutsk region, the same complex of geological and geophysical data (as on the etalon deposit) was calculated. Dynamic processing of the seismic data was carried out during the formation of data for forecasting improved reservoir properties along the Osinsky,
Preobrazhensky, and Ust-Kutsky horizons of some LAs. Attributes of the seismic wave field were calculated including AVO, acoustic impedance, and scattered-wave energy. A trend model of impedances was constructed; the results of GIS data interpretation and documents on wells testing in productive horizons were prepared. These activities resulted in forecasting and construction of forecast maps for a number of reservoir parameters. Figure 5 gives an example of comparison between the $N_{ef}$ forecast on one of the LAs using a standard technique for interpreting geological and geophysical data and interpretation based on the neural networks of the pattern created. The forecast of the northern anomaly was confirmed later by a well drilling; the deposit was discovered.

Thus, the methodology developed which is oriented on combining the principles of training and interpretability corresponds to the goals and objectives of Data Mining. It is based on the concept of patterns that reflect fragments of multidimensional relationships in data.

These patterns represent the regularities inherent in the subsamples of data, they are expressed in a human-understandable form. The search for templates was carried out by the methods not limited to a priori assumptions about the structure of the sample and the form of distribution of the values of indicators analyzed.

In the geoinformation system NeuroInformGeo, a digital geological and geophysical model was obtained for this etalon deposit and a trained neural network, which preserves the coupling coefficients between geological and field parameters and attributes of the measured geophysical fields.

This model contains digital data and an analysis tool that allows adaptive training and retraining of neural networks (based on new data or simulation of the projected wells) as well as constructing geological and geophysical models taking into account the models built at the previous stage.

3 Conclusion
The realized approach within the framework of Data Mining (discovery-driven data mining), expansion of the model and its adaptive training provides wide opportunities to refine the forecast of reservoir quality on the areas studied and poorly studied, which in turn makes it possible to clarify the financial and economic assessment of new wells, etc. Creation of new etalons (patterns) will make it possible to generalize geological and geophysical information on other territories, establish regularities affecting the efficiency of geological and exploration work, and increase efficiency of geological and exploration work.

Figure 5. Comparison between the standard methodology for interpretation of geological and geophysical data and interpretation based on the neural networks of the pattern created.
Acknowledgments
This study (research grant No № 8.2.24.2018) was supported by the Tomsk State University competitiveness improvement programme. The authors are grateful to Tatiana B. Rumyantseva from Tomsk State University for her assistance in preparing the paper.

References
[1] Gafurov O M, Gafurov D O, Gafurov A O, Pankov M V, Bitner A K, Krasilnikova N B, 2012 Patent of the Russian Federation for invention No. 2477499 “A Method for Determining Location of Production Wells During the Development of Hydrocarbon Fields”, application No. 2011112537.
[2] Gafurov O M, Gafurov D O, Gafurov A O 2011 Certificate of the Federal Service for Intellectual Property for the state registration of the computer program “The Neural Network-Based Program for Processing Seismic Acoustic and Geophysical Data in .dat Grids Format Used for Formation of Etalons and Calculation of Patterns”, No. 2011613919
[3] Gafurov O M, Gafurov D O, Gafurov A O 2011 Certificate of the Federal Service for Intellectual Property for the state registration of the database: “The Database for the Block of Automated Formation of Etalons in the Intellectual Geoinformation System “NeuroInformGeo” No 201162031
[4] Gafurov D, Gafurov O 2016 MATEC Web of Conferences 79 01086 doi: 10.1051/matecconf/20167901086
[5] Gafurov D and Gafurov O 2016 MATEC Web of Conferences 79 01055 doi: 10.1051/matecconf/20167901055
[6] Nikitin A A, Zemtsova D P, Dolinin A N 2003 Geophysics (6) 3 – 8
[7] Hunt J 1982 Geochemistry and Geology of Oil and Gas (Moscow, Peace Publ.)
[8] Pozdnyakov V A 2005 Geology, Geophysics and Development of Oil and Gas Fields (7) 30–36