INTELLIGENCE INTERFACE WITH THE DECISION MAKING EXPERT SYSTEM OF THE GIS "INTEGRO"

Eu.N.Cheremisina and M.S.Khoziainov

VNIIgeosystem, Varshavskoe shosse 8, 113105 Moscow, Russia
Fax: +7(095) 958 3711

Up-to-date prediction-diagnostic systems which make use of the expert knowledge must be organized as follows: the system's knowledge base must comprise only most general propositions, so that a practical expert might adjust the knowledge base according to his specific problem and his own concepts. Moreover, in case those general propositions do not suit him, he might change them, yet, temporarily, for a single session only. A situation where the expert's and user's functions are combined, specific demands are imposed upon the system's interface. It should have enough intelligence to guide a user, yet, at the same time it must be rather flexible to allow expert take his own way.

The computer-based INTEGRO system, which has been developed in the Geoinformatics lab of the VNIIgeosystem Institute, is intended for the integrated analysis of heterogeneous geoinformation and for solving prediction-diagnostic problems with the aid of cartographic and factographic geoinformation banks.

Consider basic levels of the interactive technology of stating and solving the prediction-diagnostic problems in geology and ecology.

The first level of the technology is an automation of the problem statement and formalization processes. The technology second level assumes the selection of optimum method for solving a formal problem by means of data complex analysis and intelligence graphic interface. This level provides the problem solution and presentation of results according to a scheme specified by a specialist - the technological scheme user. The second level also provides the analysis of attribute space, selection and implementation of an optimum method for solving the formalized problem.

System's functions are united in five groups in accordance with the main menu items. The structure of these groups is as follows:

1. Data preparation functions, including the retrieval or creation of a data set in the form of the TOP file, editing, merge of files and marking a portion of the file in the object space, sorting of data, arithmetic transformations, and visualization.

2. Problem statement: identifying the type of a formal problem; the initial formation of attribute space based on the analysis of its study; specifying the training material, if such is available.

3. Computing additional interpretation characteristics by applying numerical methods for processing geological and geophysical data.

4. Integrated data analysis consisting of the ordered set of known statistical methods allowing to make necessary changes in the data set under consideration.

5. The problem solution block as itself, including various methods of recognition of images, taxonomy, ordering and minimization of attribute space and equipped with the algorithms and rules of choosing an optimum problem solution method.

The configuration and capabilities of all blocks directly depend on the nature of the source data and problem formulation. Thus, the set of functions necessary for solving the problem posed is formed. Despite a limited number of the problem statement versions, any real problem can be represented as a series of the successive subproblems of various types.

To make the solution of those problems convenient for user and to avoid the repetition of particular steps in the problem statement procedure, the current problem solution state is saved within the framework of the interface program module.

Data analysis and problem solution algorithms implemented in this system can be used both successively - the user passes to next stage without referring to menu - and optional - in this case the user selects which data analysis methods are to be applied and he selects the problem solution method. When optional system functions are chosen, the interface is actually transformed into a logical series of hierarchical menu with the diverse visualization of data.

The successive regime relates to a complete graph describing the problem statement and solution. It provides user a full access to the intelligence content of interface. In this case the observance of system's recommendations guarantees an adequate statement of the problem and its solution by most effective method.

The combination of data format applied in the system and special tools of the interface permit one to form criterial, analog and criterial-analog models of an object under study. The criterial model represents a system of criteria favorable to the occurrence of the sought mineral resources on a given area. These criteria are presented as the properties, named criterial and distinctive in their scale values.

The property value grows with the probability of finding the sought mineral resources. It is clear that such relationship can also be inverse, though in practice such cases are rather rare.

The analog model is built based on the analogy principle, i.e. it is based on the search of objects similar to the known references objects.

Such model is built as an information classification property comprising the information about the object belonging to particular reference classes (class number). The objects which cannot be classified should have zero or blank in this property attribute. The formation of such properties is proceeded on the problem statement stage.

The initial compilation of data set is accompanied by the analysis of source information and specifying the set of
operations to be applied to data.

The problem statement stage assumes the selection of the formal problem type (ordering, division, zonation or minimization of attribute space). As far as in practice the distinctions between those statements are illegible, the problem statement also assumes the specification of additional information refining the problem statement (definition of an objective property in the ordering problem, specifying the analog model by relating particular objects to specific classes in the division and zonation problems).

Thus, even if the user meets difficulties in the problem
formulation, he may hope that the interface, first, will not allow to state a problem given a lack of data and, secondly, the interface will check up the statement correctness and change it if needed.

A special graphic editor included in the problem statement unit is designed for specifying reference objects, it allows one quickly fill and edit the reference data column.

When stating the problem, a user has to choose among available TOP properties which he would like to use in finding the problem solution.

The user can orient himself on the study map showing the completeness of information about the set of chosen properties. This allows to exclude from consideration those properties characterized by the lack of data since this fact substantially reduces the area where the solution is searched for.

The problem statement is concluded in a control unit where the validity of the problem statement is verified.

Next stage in the solution of the problem posed (when following the data processing graph installed by the designers) is the analysis of data implementing a set of widely known statistical methods. Latter are arranged so that a user can easily find necessary changes and corrections in data. The data analysis interface allows both the successive and optional proceeding of these operations.

At all the stages of data analysis the results are presented in a clear graphical form where the items are highlighted which should be solved by the user himself. At various stages of the analysis the abnormal values of varying amplitude and gaps in the matrix are highlighted. The colour scale can vary for different properties.

The opportunity for the operative correction of faulty values and removing of values of particular objects is provided.

Finally, this program block allows one to fill missed data. This operation is carried out on the basis of regression relationships between the properties, here a random component is added to calculated property values.

Such procedure permits to preserve statistical characteristics of a sample, yet, it imposes considerable restrictions: not all of the values missed can be filled through such procedure. A random component does not allow one to consider the simulated values like experimental ones, these values cannot be used in the training material in the pattern recognition problem.

The share of reliably calculated component in such value is referred to as the filling reliability index. On the basis of the map describing the filling reliability index, a user can make a decision whether to use the original or filled matrix.

The structure of the problem solution unit depends on the problem statement and can vary at similar solution finding procedures: the checking of the problem statement (including the selection of the solution method - manual or automatic) is followed by the problem solution and graphical presentation of results.

When solving the division or zonation problems where reference objects are specified, one has first to check the
representativeness of references.

One can classify rather numerous training reference material by the conformity between the scope of the property values, computed from the training objects, and the scope of values over the entire sample. After checking all the properties under study the system finally recognizes the problem as the division or zonation one and updates the user's problem statement if needed.

The analysis of representativeness is accompanied by the
construction of histograms showing the distribution of values of reference objects of all the classes and objects under study. Also the system's conclusion of the reference representativeness on a given property is displayed. Here the user can modify the TOP properties.

When stating the division problem, one has to analyze the arrangement of images in the attribute space. In case the data allows one to formulate recommendations, the actual situation is conditionally displayed and the solution method is proposed. If the user wants to take advantage of another method, he has the full right to make this.

To analyze the problem solution results, one has to build the maps of references and resulting distribution of objects over classes. One has also to build a joint map of references and division results with the opportunity to mark some objects, in order to use these marks in editing of the reference column in appropriate editor. The results of the problem solution procedures are also presented as text materials which can be printed-out.

In case no references are provided, the taxonomy methods are applied. Here the algorithm is chosen based on the assumption that the whole sample should be partitioned into uniform groups (i.e., both the measures of similarity between groups and within each of groups should be minimized, so, some average approach should be chosen).

In case the taxonomy algorithm breaks the sample into plenty of groups, only most populated groups will be visualized. Yet, it should be noted that too large number of groups points to a low efficiency of the taxonomy algorithm or evidences that the parameters do not fit the source data available.

When the ordering problem is chosen and the objective property is available, the system will apply a regression analysis, and if the objective property is absent, the heuristic algorithm will be applied based on the exclusiveness principle (the
algorithm).

When minimizing the attribute space at the lack of references, the system applies the factor analysis algorithms, and given the reference data, chooses a space guaranteeing best partitioning into images.