The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. Section #1 "Information and Communication Technologies" www.pci2008.science.az/1/26.pdf

METHODS OF MODELLING INFORMATION SYSTEM FOR PATTERN RECOGNITION AND CLASSIFICATION

Yedilkhan Amirgaliyev

Kazakh National Technical University named after K.Satpaev, Almaty, Kazakhstan
amir_ed@mail.ru

In the given work development of conceptual model of information system of recognition and classification on the basis of mathematical methods and models of pattern recognition and the classifications, intended for the analysis and processing of multivariate data is considered. The methodology of development of information system is realized with use of the object-oriented approach and carried out in language of modeling of complex program systems - the unified language of modeling UML. The considered concept of information system construction can be used in different subject areas, as the modern concept of development of similar information systems. For realization of design decisions it is used CASE-means Rational Rose.

In a described method precedents of use, static modeling, and diagrams events sequence, which meet in several methods, are combined. The applied notation is based on UML. During modeling precedents the functional requirement to system in terms of actors and precedents is defined. The static model offers a static sight at information aspects of system. The class is defined in terms of the attributes and mutual relations with other classes. Result of dynamic modeling is the dynamic sight at system. The precedents formulated earlier with the purpose to show interaction of the objects participating in each it are specified. Diagrams of cooperation and the sequences reflecting cooperation of objects in each precedent are developed. Aspects of system depending on a condition it is described by means of diagrams conditions, and for each object the diagram is made.

As a mathematical ware of developed system the mathematical device of the decision of a problem of recognition and classification, a problem group classifications and optimization models using various kinds of quality functional are considered.

Statement of a classification problem Z_k**.** Let have given initial information *I*, {*S*} – set of admissible objects and each object $S_i \in \{S\}$, $i=1,..., m$ is characterized by *n* - measured vector, which coordinates call as the attributes taken from the alphabet of attributes, i.e.

$S_i=(\alpha_{i1}, \alpha_{i2},...., \alpha_{in})$, $i=1,..., m$, n – number of attributes, $\alpha_{ij} \in M_j$.

It is required to construct algorithm of classification *A* for sets {*I*(*S*) , *S*}, classifying initial set of objects *S* on a number of not crossed classes (clusters), K_j, $j=1,..., l$, so that the objects belonging to one class (cluster) were similar (relatives) while the objects belonging various classes, were in some certain sense unlike (removed):

$$A\{I(S), S\} = \bigcup_{j=1}^{l} K_j, K_i \bigcap K_j = \varnothing,$$

if $i \neq j$, $K_i \neq \varnothing$, $i, j=1, 2,..., l$.

Cluster construction can be considered as a problem of pattern recognition without the teacher, considering, that on the given set of objects, as a rule, there is no information, concerning numbers of classes and structures of classes.

At construction of algorithms by a principle of a minimum of distance, clusters search and definition of patterns are questions of primary importance. At construction of such algorithms two approaches, as a rule, are used. One of them - heuristic, and in its basis lays intuition and experience. The second approach provides minimization or maximization of some chosen parameter of classification quality.

In models group classifications the metrics used in space of classifications is entered. We will consider a problem of synthesis of group classification Z_C. Let $A_1,..., A_m \in \{A\}$ is an initial set of algorithms of the decision of a problem of classification Z_k for set of objects $M = \{S_1,...,S_n\}$.

Result of application of algorithms A_i to set $(M, J(M))$ are classifications $K_i(M) \in \Re(M)$, $\Re(M)$ – space of classifications of final set of objects M, which elements are separate classifications. Let defined the metrics $d(K',K'')$ in $\Re(M)$ and $\varphi(K) = \sum_{i=1}^{m} d(K,K_i),\ K_i = K_i(M),\ K \in \Re(M)$.

Then the primary task of group synthesis (group classifications) Z_C consists in the following: find the classification $K^*(M) \in \Re(M)$, minimizing the functional $\varphi(K)$, i.e. $\varphi(K^*) = \min \varphi(K),\ K \in \Re(M)$.

We will consider space of classifications $\Re(M)$ of set M. It is known, that on any classification $K \in \Re(M)$, it is possible to construct the binary attitude corresponding it R, which is the attitude of equivalence on set M.

Let's consider standard representation of classifications $K(M)$ from $\Re(M)$ in the form of final set of classes $K_i(M)$ – subsets of set M, i.e. $K(M) = \{K_1(M), K_2(M),...,K_l(M)\}$ that is for the description of classification there is enough transfer of objects numbers which have got in each classes. And, as it will be shown below, the way of numbering of classes can be any for received any classification, and does not influence result of their comparison. At such representation of classifications for the decision of the task Z_C it is necessary to have the metrics in $\Re(M)$, which full enough would reflect distances really existing in given space, and to investigate properties of the space $\Re(M)$, being structure.

We will designate through $K^l(M)$ set of classifications M on l, $1 \le l \le n$ classes, i.e. $\bigcup_{l=1}^{n} K^l(M) = \Re(M)$. Let $K_n(l)$ - any classification from $K^l(M)$. We will set representation $d : K(M) \times K(M) \to Z$ by means of the following formula:

$$d(K_n{}^u(l_u), K_n{}^v(l_v)) = 2n - \sum_{j=1}^{l_u} \max_{l \le i \le l_v} \left\{ \left| K^u_{n,j}(l_u) \cap K^v_{n,i}(l_v) \right| \right\} -$$

$$- \sum_{i=1}^{l_v} \max_{l \le j \le l_u} \left\{ \left| K^v_{n,i}(l_v) \cap K^u_{n,j}(l_u) \right| \right\},$$

where $$K^t{}_n(l_t) = \left\{ K^t{}_{n,1}(l_t),...,K^t{}_{n,l_t}(l_t) \right\},\ 1 \le l_t \le n,\ t \in \{u,v\}.$$

The entered metrics has properties of the metrics [3].

Conceptual model of information system of recognition and classification. The conceptual scheme of the developed information system of recognition and classification consists from several modules.

We will briefly describe functional purposes of the subsystems, which are a part of system: Subsystem ***Help*** is presented as information system. Shows the information on functionalities, both separate subsystems, and system as a whole; ***Management*** - the subsystem represents the program interface of control process of projected system; the ***Database*** – set of some tables for a data storage necessary for system; the subsystem of ***preliminary processing*** is intended for preliminary processing initial data: definitions of the blank data and definition of optimum subsystems of attributes in the description of objects; definition of informative attributes.

Models and algorithms of classification work as algorithms of classification making a base set for system; **Models of group synthesis** represent work of algorithms of the group synthesis realized within the limits of system; **Visualization of results** represents results of work, both subsystems, and systems as a whole for users in the necessary format; **Analysis and estimation of results** intended for the analysis and an estimation of results on the basis of parameters of the chosen kinds of qualities functional and represents a basis for development of the recommendation about use of various computing schemes; subsystem GIS modeling represents opportunities of use of the developed methods, algorithms of group synthesis within the limits of geoinformation modeling.

Within the limits of information system, depending on statements of the user tasks, it is possible to carry out optimization procedures, using concrete kind qualities functional. The scheme of carrying out of optimization procedures realizes a subsystem the *Analysis and estimation* (Figure 1).

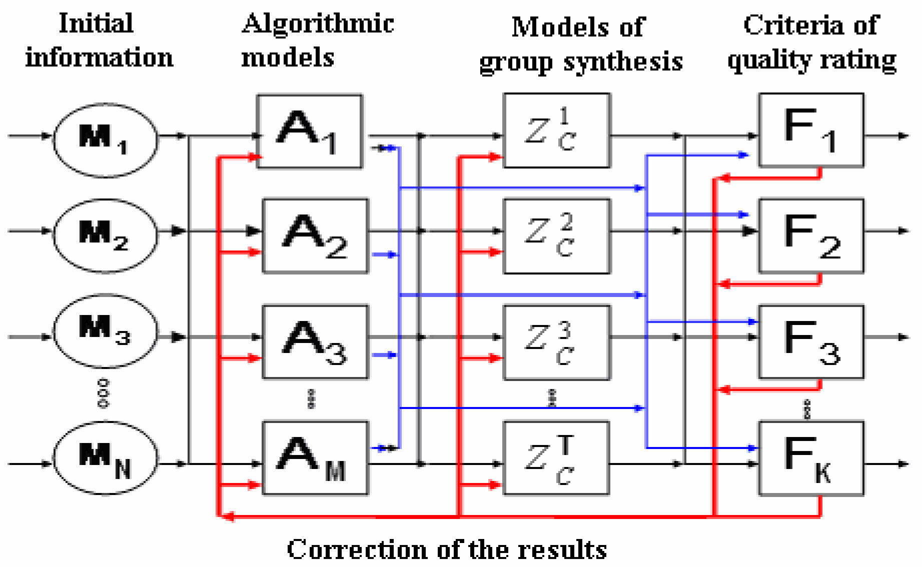

Figure 1. *Scheme optimization procedures (two-level model).*

Here: M_N – set of initial data, A_M – algorithms of classification, Z_C^T – models of group synthesis, F_K – quality functionals.

Within the limits of the given scheme following models of optimization (various combinations of algorithms, models of group synthesis functionals of quality are applied) are possible: single-level model M_1 $\{(M_k, A_j, F_t): k=1, \ldots, N; j=1,\ldots, M; t=1,\ldots, N.\}$; two-level model M_2 $\{(M_k, A_j, Z_c^i, F_t): k=1, \ldots, N; j=1,\ldots, M; i=1,\ldots, T; t=1,\ldots, K.\}$.

During modeling system following kinds of diagrams (notations) UML are used: *diagrams of precedents* represents functionality and behavior of developed system, allows to define system requirements to define objects operating in system and the main tasks, which are carried out by these objects (representation of scripts); *diagrams of classes* provide static design representation of system; *collaboration diagrams* describe interaction of objects, abstracting from sequence of data transmission, accepted and transferred messages of concrete object and types of all these messages are reflected; *sequence diagram* allow to define sequence of message transfer between objects, shows a stream of messages; *state diagram* are intended for display of state of the system objects having complex model of behavior.

For management of functioning of system it is created, a so-called ***session of processing***. The ***session of processing*** – choice of computing process and formation of set of input and output data for the certain subsystem and process of data processing. *Input data* - initial data, the list of algorithms and parameters. *Initial data* – subset of input data of the subsystem, which is the general for all algorithms of this subsystem. The structure of initial data is regulated by the subsystem. *The list of algorithms* - subset of algorithms of the subsystem, which execute processing in the given session. *Parameters* – a subset of input data of the subsystem, which is not the general for its all algorithms. Parameters are grouped on algorithms. Each algorithm declares the set of parameters and their structures. Output data - set of return codes and results. *Return code* - the numerical value designating the status of the termination of algorithm.

Create a session – creation of a new session. *The precondition* - system have not started or the session is in state *new*. *The basic stream of events*: the user, starting the program, or, using elements of management of the user interface of the started program, creates a new user session. *The post condition* – system is in state – New session.

Also within the limits of the session precedents – **open a session, work with the session** are considered.

Also under the specified scheme other precedents of the session of processing specified on the diagram of precedents are described.

Thus, in the given work we showed some aspects of modeling of information system of recognition and classification with application of the unified language of modeling. For realization of project decisions are used toolkit CASE-of means Rational Rose.

The developed information system is applied to the decision of real applied tasks from hydrogeology area, ecological monitoring with use of data of remote sounding and geology.

Literature

1. G. Buch Object-oriented analysis and designing. – M.: «Publishing house Binom», 1999.
2. W. Boggs. Rational Rose & UML. – M.: «Publishing house Lori», 1999, - 480 p.
3. M. Aidarkhanov, E.Amirgaliyev. Algorithmic basis of construction of classification systems. - Almaty, 1998. – 100 p.
4. A. Vendrov. Designing of the software. – M.: 2000, - 290 p.