Computer Science Journal of Moldova, vol.**7**, no.1(19), 1999

Intelligent tecnology of information analysis and quantitative evaluation of condition with preliminary and inaccurate data of current observations

Yu. Savva

Abstract

The article presents an approach to design of data processing systems of ecological monitoring based on modern information technologies: data mining, genetic algorithms and geoinformatics.

1 Introduction

Quantitative evaluation and forecasting of ecological condition is of the utmost importance. However a range of factors hampers the solution of this problem:

- large dimensionality, that does not depend on the size of the research region and its ecological system;
- absence of regular, accurate and reliable parameter values, characterizing the condition of an ecosystem;
- lack of our knowledge about the principles of interaction of certain components of ecosystems and its reaction on different anthropogenic influences.

One of the main properties of ecosystems is their ability to self-restoration after various external influences. However this ability is

not boundless. Regeneration process is possible only when changes in certain quantitative indices will occur within the limits of a certain range of values. These values exceed the limits of compensation ranges for destructive changes occur in ecosystems, that impede the process of self-restoration, and their quantitative qualitative accumulation can lead to degradation and ever destruction of ecosystems.

In this connection the problem of determination of integral indices for quantitative evaluation of ecosystem condition arises. Also it is necessary to detect the number of indices that influence the dynamics of ecosystem development the most.

2 Integral criteria of quantitative evaluation of ecosystem condition

The condition of an ecological system can be characterized by a great number of indices. Measurement of these indices is not regular and measuring instruments have a certain error of measurement. That is why quantitative evaluation of ecosystem condition and forecasting of its development that can be obtained by means of statistical methods and calculus theory have very little to do with reality, because their application in these conditions is not correct.

The use of dimensionless criteria for evaluation of condition and dynamics of change of ecological situation in a region is substantiated in our work [1] and the approach of comparison of ecological situation in two regions with the use of these criteria is offered.

These criteria are:

1. Relative change of ecological situation in a region (in an object), defined as:

$$Q\left(t_{m_1}, t_{m_2}\right) = \cos\left(\varphi\left(t_{m_1}, t_{m_2}\right)\right) = \frac{\left(R'\left(t_{m_1}\right), R'\left(t_{m_2}\right)\right)}{\left\|R'\left(t_{m_1}\right)\right\| \left\|R'\left(t_{m_2}\right)\right\|},$$

2. Velocity of change in ecological situation in a region, defined as:

$$V\left(t_{m_1}, t_{m_2}\right) = \frac{Q\left(t_{m_1}, t_{m_2}\right)}{\Delta t} = \frac{Q\left(t_{m_1}, t_{m_2}\right)}{t_{m_2} - t_{m_1}}.$$

Here, $R'(t_m)$ is the information frame of ecological condition in a region, defined in various periods of time t_m, that represents the complex of graphically presented data spatially distributed on the map and characterizing the ecological situation in the research region.

Forming of information frame of ecological situation of a region is carried out on the base of parameter vector $D = \left\{ d_k \mid k = \overline{1, K} \right\}$ characterizing the condition and functional value of observed object. Meanwhile, determination of this vector makes up the greatest difficulty because of the lack of information and low data authenticity of current observations upon the objects. That is why to build the representation $f : D \times R \to R'$ it is necessary to take into account the fact that R is an exactly definite graph in the form of net interconnected objects; D is accurately defined subset. Then the resulting information frame R' will represented an accurately defined graph from the point of view of Berge and Kenig.

As a result of this the evaluation of relative change of ecological situation in a region should be scaled according to the maximum vector values D (maximum permissible concentration (MPC), sanitary norms, etc.). Then the evaluation of environmental safety in a region can be determined as a general Hamming's distance $r\left(D, \hat{D}\right)$, where $\hat{D} = \left\{ \hat{d}_k \right\}$ is a vector of maximum values of parameters of vector D.

Using the given approach it is possible to estimate environmental safety of regions in spite of the absence of full and authentic information about all parameters, characterizing ecosystem condition, and also their classification. Besides, it is possible to model ecosystem development and visualize their conditions by geoinformation systems (GIS).

3 Detection of important indices for the evaluation of ecosystem condition

As it was mentioned above, ecosystem condition can be fully described by a great number of indices $D = \left\{ d_k \mid k = \overline{1, K} \right\}$ that have numerical and non-numerical nature. It is obvious that use of all of these

indices will lead to the necessity of solution the problem with large dimensionally. Consequently it is necessary to choose from all the indices D characterizing the ecosystem condition and influencing its dynamics the most, a certain subset Ω. This subset $\Omega \in D$, $\Omega = \left\{ \omega_\psi \mid \psi = \overline{1, \Psi},\ \Psi < K \right\}$ represents a range of names of the most informative indices.

Determination of elements of subset Ω requires the analysis of ecological information in order to detect the concealed phenomena in a state of important features, correlations and complex relations in multidimensional data.

Consider the number of indices characterizing the ecosystem condition as a certain combination of defining the quality of an individual. Then values of parameters identifying one or the other ecosystem condition can be named "genes". Under these assumptions the process of artificial change of elements of subset Ω will remind of controlled by a person evolution of population of individuals through the influence on a given range of "genes". Here we can observe the 3 main mechanisms of evolution:

1) mutations — accidental changes of "genes" in some individuals in populations;

2) recombinations — production of new individuals with the help of mixing "geneous" ranges of selected individuals in populations and

3) selection of the strongest "genes" that give the most accurate descriptions on ecosystems.

The result of change of generations and selection the range of elements of subset Ω can not be improved under the achieved level of knowledge about ecosystems and mechanisms of their functioning.

Thus the problem of definition of elements of subset Ω comes down to a definition of the criteria of "gene" selection.

Algorithmization and automatization of this procedure not only speed up the procedure itself but, which is more important, allow to

search and reveal in the input arrays of data available but not always evident associations between various factors improving the objectivity of a choice. These associations can be represented as rules in their left parts. They allow to determine frequency of appearance of one element in input arrays of data with other elements. In particular one of the rules can be expressed by the conditional relation: "IF–THEN" (implication), which means: IF some condition is performed, THEN a certain action follows. For example, IF heavy rains we in the upper reaches of the river, THEN in a certain period of time there will be floods in the low reaches. Numbers associated with these rules determine their forecasting power.

Let's bring binary set $A = \|a_{ji}\|$, $j = 1, 2, \ldots, n$; $i = 1, 2, \ldots, m$, in correspondence with full set of information elements D, characterizing the ecosystem condition, where each element a_{ji} can be either 0 or 1. Here $a_{ji} = 1$ corresponds to some event X, and $a_{ji} = 0$ to event $\overline{X}$ occurred in the ecosystem for an analyzed period of time.

According to the stated above X is an event corresponding to $a_{ji} = 1$ hence the event k occurred in the period of time j. With the help of numbers Z1, Z2 and Z3 it is possible to determine if there is a connection between the events X_i and X_k or continuing our example determine the forecasting power of the rule IF–THEN.

Let's define the entity of numbers Z1, Z2 and Z3 and list the formulas for their calculation.

Number Z1 — forecasting of a rule

$$\mathrm{Z1} = \frac{\sum\limits_{j=1, i\neq k}^{n} (a_{ji} \& a_{jk})}{\sum\limits_{j=1}^{n} a_{ji}},$$

defines how often information elements, corresponding to events X_i and X_k appear together as a part of the number of records, corresponding to the event X_i.

Number Z2 — prevalence of a rule

$$\mathrm{Z2} = \frac{\sum\limits_{j=1,i\neq k}^{n} (a_{ji} \& a_{jk})}{n},$$

shows how often elements, corresponding to events X_i and X_k appear together in the array of initial data as a part of total number of records in this array.

Number Z3 — expected value of forecasting

$$\mathrm{Z3} = \frac{\sum\limits_{j=1,i\neq k}^{n} \neg\,(a_{ji} \circ a_{jk})}{n - \sum\limits_{j=1}^{n} a_{ji}},$$

where $\circ$ is a Webb's function.

This number shows the frequency of appearing of elements in the right part of a rule. It expresses the forecasting that exists under the absence of connection between the elements.

Information got by these numbers allows making important conclusions. E.g., analyzing 620 indices characterizing the ecosystem of Oka-river in Orel region (Russia) for the last few years we could conclude that only 17 really influence the change in ecological situation.

Consequently, while elaborating the mathematical model and substantiating the solutions in the field of ecology, in the first place it is necessary to take into account the indicated 17 factors.

4 Algorithm of solution of a problem

The result of algorithm run are the arrays MZ1, MZ2 and MZ3, whose elements are correspondingly $z1_{ik}, z2_{ik}, z3_{ik} \mid i,k = 1,2,\ldots,m;\ i \neq k$; they are used to detect the connections between the all possible pairs of events.

Variable is the counter of events X_i.

Step 1. $i := 1$.

Step 2. For each $k = 1, 2, \ldots, m$ under condition, that $k \neq i$, perform:

$$z1_{ik} := \sum_{j=1}^{n} (a_{ji} \cdot a_{jk});$$

for each j from 1 till n:

if $a_{ji} = 1$ and $a_{jk} = 1$, then $z3_{ik} := z3_{ik} + 1$;

$$s := \sum_{j=1}^{n} a_{ji};$$

$$z2_{ik} = z1_{ik} \,/\, n;$$

$$z1_{ik} = \begin{cases} z1_{ik} \,/\, s, & \text{if } s \neq 0, \\ -1, & \text{otherwise} \end{cases};$$

$$z3_{ik} = \begin{cases} z3_{ik} \,/\, (n - s), & \text{if } s \neq n, \\ -1, & \text{otherwise} \end{cases};$$

Step 3. Increase the value $i := i + 1$. If $i \leq m$, then repeat all the procedures beginning from step 2, otherwise the algorithm run is over.

Conclusions

Practical application of the given approach to the analysis of ecological situation in Orel region (Oka-river) showed that:

1. Introduced for integrated evaluation of changes in ecological situation, integral indices "relative change in ecological situation" and "velocity of change in ecological situation" can be used for the analysis of trends in the development of this situation and also for comparison of these trends in various regions.

2. Based on methods of intellectual analysis of data, means of representation and interpretation of multidimensional data allow to

reveal possible but not evident associations between indices characterizing an ecosystem condition, to find the objective clusterization, pick out significant ("critical") indices for generation of models.

3. Elaborated algorithms and "system of quantitative analysis of ecological information" software showed high efficiency of work and significance of the obtained results.

References

[1] Y.B. Savva. Integral Criteria of Estimation the Change in Ecological Situation// 1-st Int. Conf. on Problems of Ecology and Vital Activity Protection. Tula, 1997, pp.438–440.

Yuri B. Savva,
Orel State Technical University
Department of Information Systems
29, Naugorskoe shosse,
Orel, 302020 Russia
phone: +7+0862+413295
fax: +7+0862+416684
e-mail: *yurisa@ostu.ru*

Received December 4, 1998