Intellectual technology of detection of anomalies in the aquatoria ecosystems of the Sevastopol on the basis of data clustering

A. Skatkov[1], *A.* Bryukhovetskiy[1], and *D.* Moiseev[1,2,*]

[1]Federal State Autonomous Educational Institution of Higher Education «Sevastopol State University», Russian Federation, 299053, Sevastopol
[2]Black Sea Higher Navy Order of the Red Star Academy by P.S.Nakhimov, Federal State Owned Military Educational Institution of the Higher Professional Education, Russian Federation, 299028, Sevastopol

Abstract. The main features associated with the development of intelligent technology for detecting anomalies of ecosystems in the waters of the city of Sevastopol are considered. An approach is proposed, the feature of which is to ensure continuous monitoring of key environmental indicators presented in the form of heterogeneous information flows: hydrometeorological information, data on the level of pollution and air composition, soil, environmental monitoring, monitoring of maximum permissible emissions of harmful substances in order to detect changes in the state of data flow monitoring. The proposed method for the detection of anomalies of ecosystems of the water area is based on data clustering. We consider typical operations on clusters and main metrics based on the Kullback information measure.

1 Introduction

With the growth of cities, the development of production, the technogenic transformation of the environment is becoming too global. This paper is devoted to the study of the processes of detection of anomalies of ecosystems of the water area of the city of Sevastopol on the basis of data clustering. Environmental monitoring and control is a set of measures to identify and assess the sources and levels of contamination of natural objects with harmful substances as a result of discharges or emissions of these substances into the environment by nature users, as well as due to natural formation and accumulation in environmental objects, including through chemical and biochemical transformation of natural and man-made substances into compounds with harmful properties [1]. The lack of comprehensive monitoring studies of the background state of the marine environment on the seashore of Sevastopol, including oil and phenolic pollution, control of which is especially needed in the waters adjacent to the infrastructure of the cargo-passenger and military fleets, does not allow an adequate assessment of the state of ecosystems and the future development of

* Corresponding author: dmitriymoiseev@mail.ru

environmental situations [2]. Therefore, the development of methods for monitoring the detection of ecosystems anomalies in the waters of the city of Sevastopol is an urgent task.

2 Review of literary sources

Currently, the process of quickly identifying anomalies of environmental monitoring data and critical infrastructure is a complex, time-consuming and difficult to formalize task. This is due to the fact that, as a rule, each instance of monitored monitoring data has a complex multidimensional structure and may include many variables and parameters[3]. The combined use of operational monitoring tools, simulation modeling and probabilistic models allows us to predict the dynamics of changes in the state of the ecosystem, warn of possible anomalies and preventively take corrective actions, thereby preventing the occurrence of emergency situations[4]. To date, the tasks of system monitoring include the tasks of monitoring, managing and operational forecasting changes in complex systems. The use of information technologies that use high-performance distributed high-loaded computing systems allows us to quickly and cost-effectively solve many applied problems of monitoring, analyzing and forecasting processes occurring in complex critical systems, including natural systems. One of the software and methodological implementations used for the processes of spatial and geospatial modeling of structure, interconnections and dynamics in hydro and oceanology are geographic information systems (GIS)[5]. Using the model of distributed or cloud computing using an agent-based approach used to process Big Data will allow you to make the transition to a qualitatively new knowledge, optimize the processes occurring in large-scale scientific research, and solve the problem of analysis and representative visualization of measurement slices in multidimensional databases [6-8], to analyze the cyclical nature of changes in ecological and geological changes, while using statistics for more than 100 years of environmental monitoring.

In order to develop well-known approaches to solving these problems, a new scientific approach is proposed, which largely overcomes the requirements for implementing known methods for monitoring the state of ecosystems, in particular:

- large amounts of a priori information for setting the parameters of the monitoring system;
- deciding on the state of the ecosystem as a result of processing a large amount of current data;
- it is difficult to implement a real-time data processing strategy.

3 Description of the method

The proposed approach is based on data clustering, which is widely used in image recognition and data mining tasks. Of particular interest to the methods of data analysis arose in connection with the development of tools for collecting and storing large amounts of data from numerous sensors. With the emergence of new technologies “Internet Things” and “Big Data”, specialists from different areas of human activity are faced with the task of processing large amounts of data in real time and with a high level of confidence [9 - 11].

This article proposes an anomaly detection model based on dynamic clustering, which examines the following situations that change the structure of clusters:

- formation of new clusters;
- cluster merging;
- cluster splitting;
- disappearance of clusters;
- drift of cluster centers.

The first four types of structural changes are abrupt (jump-like) changes in the cluster structure. In many applications, such drastic changes are associated with deficiencies in the behavior of the observed system, and detecting such changes as early as possible helps to avoid various undesirable consequences. The fifth type of structural change is continuous (and usually minor) in nature, which is not always easy to detect, but which is also of great importance in practice.

The identification of dynamic changes and the methods of corresponding corrections of the cluster structure will be considered separately for each of the listed types of these changes.

1. Identification of new clusters. After the classification of new objects (with a cluster structure known from the previous time window), it is necessary to determine at the time point whether the appearance of new objects has led to the appearance of new clusters. It seems that the announcement of the emergence of new clusters requires the following conditions:

• the presence of new objects with small degrees of belonging to all existing clusters ("free" objects);

• the sufficiently large number of such objects, significantly exceeding the number of existing clusters;

• the compactness of these objects: they must form a compact group.

2. Detection of merging clusters. By analogy with the previous one, we can formulate three conditions, the fulfillment of which allows declaring two clusters to merge:

• the presence of objects with high degrees of belonging simultaneously for two clusters;

• the sufficiently large number of such objects;

• the proximity of the centers of two corresponding clusters.

3. Identification of fissionable clusters. By analogy with the merging of clusters, the reason for their splitting is the assignment to any of the already existing clusters of a large number of new objects, which can lead to heterogeneity of its internal structure. A more subtle criterion for such a splitting is the multiextremality of the histograms of features constructed for objects belonging to similar clusters.

4. Detection of disappearing clusters is a fairly simple operation: the cluster is declared disappeared if no object from the last time window was assigned to it.

5. Detection of cluster centers drift. Over time, new registered objects can cause slow changes in the positions of the cluster centers.

The main metrics of clusters (Kullback divergence is used as a distance):

i $(i=1,K)$ – number of clusters,

Mi – power of i-th cluster,

Ci – center of the i-th cluster,

D_{ij} – distance between centers of clusters C_i and C_j ,

$\overline{dip} = \frac{1}{Mi}\sum_{p=1}^{Mi} dip$ – the average distance between the center of the i-th cluster C_i and the value of the sign $\boldsymbol{x}_p \in C_i$,

ρ_l – threshold values for the Kullback distance, by which the information states of objects belonging to specified clusters are distinguished.

To assess the homogeneity of two random samples, we use the concept of Kullback-Leibler divergence (J-effect). Calculation of Kullback-Leibler divergence was performed in accordance with the formula [12]:

$$J = \sum_{i=1}^{k}\left(\frac{f_i}{m_1} - \frac{q_i}{m_2}\right)\ln\frac{f_i m_2}{q_i m_1} \tag{1}$$

where *fi, qi* – the number of hits in the i-th interval of examples of compared samples, *m1, m2* – the number of both samples.

In view of the above, the general algorithm can be described as follows:

1) monitoring is organized for the set of objects under study by fixing the signs of these objects in the time windows; objects captured in the first time window are clustered using a J-metric. In this case, the initial number of clusters is selected iteratively or is given from any a priori considerations;

2) in the second and subsequent time windows, a check of the dynamic changes of the cluster structure is carried out (identification of new clusters, merging, splitting, disappearing clusters, as well as drift of the centers of existing clusters);

3) if necessary, the cluster structure is corrected in accordance with the identified changes;

4) after detecting changes and a corresponding correction of the state of the current cluster structure, the ecosystem condition can be predicted.

4 Conclusion

The development of intelligent technology for detecting anomalies of ecosystems in the water area of the city of Sevastopol, based on the use of new approaches and methods, will lead to an increase in the validity, reliability and efficiency of decision support processes for evaluating the probability of accepting hypotheses about the presence of anomalous values, taking into account errors of the first and second types.

Adaptive decision-making methods in the face of uncertainty will eliminate the shortcomings and limitations inherent in classical approaches in the case of noisy data and incomplete information. On the basis of big data technology and a special modeling stand, the quality of evaluating decisions is increased.

The practical significance of the work results will lead to a decrease in the level of negative impact of natural and anthropogenic factors on the state of ecosystems of the water area of the city of Sevastopol, to a reduction in the negative anthropogenic impact on the environment, and also have a positive impact on the health and living conditions of the population.

This work was carried out with the partial support of the Russian Foundation for Basic Research (grant No. 18-47-920007 \ 18).

References

1. The current state of pollution of the waters of the Black Sea / Ed. A.I. Simonov, A.I. Ryabinina // Hydrometeorology and hydrochemistry of the seas. t.IV. Black Sea. v.3.– Sevastopol: ECOSI-Hydrophysics, 1996. - 230 p.
2. Ovsyany EI, Romanov A.S., Minkovskaya R.Ya., Krasnovid I.I., Ozyumenko B.A., Tsymbal I.M. The main sources of pollution of the marine environment of the Sevastopol region // Environmental safety of the coastal and shelf zones and the integrated use of shelf resources. Sevastopol: ECOSY-Hydrophysics, 2001.– P. 138-152.
3. Bryukhovetsky A.A., Skatkov A.V., Shishkin Yu.E. Simulation of anomaly detection processes in complexly structured monitoring data // Environmental control systems. 2017. No. 9 (29). Pp. 45–49.
4. Novikova A.M., Averyanova E.A. Application of GIS-technologies for solving complex problems of spatial modeling in oceanology and ecology // Environmental problems of the Azov-Black Sea region and integrated management of biological resources: materials scientific-practical. young conf. Sevastopol, 2016. pp. 200–203.

5. Bondur V.G. Aerospace monitoring of oil and gas facilities. M .: Scientific world, 2012. 558 p.
6. Scheduling in distributed systems: A cloud computing perspective / L.F. Bittencourt, A. Goldman, R.M. Madeira [et al.] Computer Science Review. 2018. Vol. 30. P. 31–54.
7. Monsalve S.A., Carballeira F.G., Calderon A. A heterogeneous mobile cloud computing model for hybrid clouds // Future Generation Computer Systems. 2018. Vol. 87. P. 651–666.
8. Vagin V.N., Golovina E.Yu., Zagoryanskaya A.A., Fomina M.V. Reliable and plausible inference in intelligent systems. - M .: Fizmatlit, 2008. 712 p.
9. Barsegyan A.A., Kholod I.I., Tess M.D. and others. Analysis of data and processes. - SPb: BHV-Petersburg, 2009. 512 p.
10. Data mining in super-large databases / V. Ganty, J. Gerke, R. Ramakrishnan // Open Systems, No. 9-10, 1999.
11. Gimarov V.A., Dli M.I., Kruglov V.V. Temporal variability of images // Vestnik MEI. 2003. № 2. S. 91-95.
12. Kulbak S. Information Theory and Statistics. - M .: Science, 1967. - 408 p.