Data mining technologies in a distributed medical information system

V E Bolnokin¹, D I Mutin², E I Mutina², V G Vyskub³ and O Ja Kravets⁴

¹ Mechanical Engineering Research Institute of the Russian Academy of Sciences, 4, Malý Haritónievsky all., Moscow, 101000, Russian Federation
² Moscow State Technological University “STANKIN”, 3a, Vdovkovsky all., Moscow, 127055, Russian Federation
³ Scientific Research Institute – Federal Research Centre for Projects Evaluation and Consulting Services, 13/1, Antonova-Ovseenko str., Moscow, 123317, Russian Federation
⁴ Voronezh State Technical University, 14 Moscow ave., Voronezh, 394026, Russian Federation

E-mail: e.mutina@mail.ru

Abstract. Proposed a method for solving the problem of identifying hidden relationships in hard-to-structure data that have an implicit character is considered using information mining. Proposed decision trees, the effectiveness of which is illustrated by a specific example. The use of OLAP analysis systems on data presented using in the form of a real or virtual hypercube’s information is an effective tool for the effectiveness of the management for medical monitoring

1. Introduction
One of the main elements of a distributed information system is difficult to structure data, the analysis of which is necessary to improve the efficiency of this system, but it is often difficult to implement.

Let us consider the decision support system (DSS) [1, 2], which is a system for the diagnosis and differential diagnosis of diseases.

All medical problems that have an implicit character are solved by explicit methods with accuracy and convenience, which are completely insufficient for widespread practical use in specific problems of forecasting and decision-making [3]. Implicit problems of medicine have become an ideal field for intellectual data mining (IDM) [4].

Decision support systems are complex software and hardware complexes designed to help decision makers manage complex objects and processes of various nature [5]. The information based on which the decision is made must be reliable, complete, consistent and adequate.

As a rule, when organizing DSS, so-called data warehouses are used, which perform the functions of preliminary preparation and storage of data for DSS based on information from the organization's automated management system, as well as information from third-party sources [6].

The approach to data search and retrieval, which is based on the idea that there is a relationship between the frequency of requests and the degree of aggregation of the data with which the requests operate (the more aggregated the data, the more often the request is executed), is called On-line Analytical Processing (OLAP) [7]. Basic in OLAP is the concept of a multidimensional data cube, in the cells of which the analyzed (numeric) data is stored.
2. Multidimensional data mining

IAD is understood as a decision support process based on the search for hidden patterns (information patterns) in data [8,9]. Operational analytical processing and intelligent data analysis are two integral parts of this decision support process. The multidimensional IDM tool (figure 1) finds patterns not only in detailed data, but also in aggregated data with varying degrees of generalization.

![Figure 1. Architecture of the multidimensional IDM system.](image)

A study of IDM methods in medical information systems (MIS) has shown that decision trees are the most effective for solving medical problems [10].

From the point of view of the field under consideration, CART (Classification and Regression Tree) is the most preferred among algorithms that implement decision trees, since it allows solving not only classification problems, but also regression problems [11].

3. CART mathematical tool

In the CART algorithm, each node in the decision tree has two children. At each step of tree construction, the rule generated at the node divides the specified set of examples (training set) into two parts – the part in which the rule is executed and the part in which the rule is not executed. To select the optimal rule, the partitioning quality evaluation function is used, which is based on the intuitive idea of reducing impurity (uncertainty) in a node. In the CART algorithm, the idea of “impurity” is formalized in [11, 12, 13].

If the dataset $T$ contains data from $n$ classes, then the Gini index is defined as:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2$$  \hspace{1cm} (1)

where $p_i$ is the probability (relative frequency) of class $i$ in $T$.

If the set $T$ is divided into two parts $T_1$ and $T_2$ with the number of examples in each $N_1$ and $N_2$, respectively, the partition quality indicator is:

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2)$$  \hspace{1cm} (2)
The best partition is considered to be the one for which $Gini_{split}(T)$ is minimal. We denote $N$ – the number of examples in the ancestor node, $L$, $R$ – the number of examples in the left and right descendants, respectively, and $l_i$ and $r_i$ – the number of instances of the i-th class in the left and right descendants. Then the quality of partitioning is evaluated:

$$Gini_{split} \rightarrow \min$$

where

$$Gini_{split} = \frac{L}{N} \left( 1 - \sum_{i=1}^{n} \left( \frac{l_i}{L} \right)^2 \right) + \frac{R}{N} \left( 1 - \sum_{i=1}^{n} \left( \frac{r_i}{R} \right)^2 \right).$$

Since multiplication by a constant does not play a role in minimization, the following transformations of the selection criterion are possible:

$$Gini_{split} = L - \frac{1}{L} \sum_{i=1}^{n} l_i^2 + R - \frac{1}{R} \sum_{i=1}^{n} r_i^2$$  \hspace{1cm} (4)

$$Gini_{split} = N - \left( \frac{1}{L} \sum_{i=1}^{n} l_i^2 + \frac{1}{R} \sum_{i=1}^{n} r_i^2 \right)$$  \hspace{1cm} (5)

$$Gini_{split} \rightarrow \min$$  \hspace{1cm} (6)

$$\hat{G}_{split} \rightarrow \max$$  \hspace{1cm} (7)

where

$$\hat{G}_{split} = \frac{1}{L} \sum_{i=1}^{n} l_i^2 + \frac{1}{R} \sum_{i=1}^{n} r_i^2.$$

Thus, the best partition is the one for which the value of $\hat{G}_{split}$ is maximal.

4. Application of data mining in a distributed medical information system

The necessary information is extracted from the databases of information subsystems and placed in storage. The logical model of the data warehouse is shown in figure 2.

To create and manage a data warehouse, the Microsoft SQL Server DBMS is used due to the simpler installation procedure and the greatest applicability and preference in MIS [1]. In addition, the main advantage of this DBMS is decision support services (Microsoft Decision Support Services) - these are tools that allow you to make publicly available the capabilities of OLAP analysis and information stored in the data warehouse.

Microsoft SQL Server Analysis Manager analytical services administration tool contains an implementation of algorithms for some IDM methods, in particular, the algorithm for building decision trees (Microsoft Decision Trees) [14].

As a result, it is easy to see tight integration between the data warehouse and decision trees. OLAP analysis using various operations allows you to collect the necessary data, store it, and make cross-sections of hyper cubes. Decision trees based on the information placed in them help to identify the dependence of changes in observation results on certain indicators. Based on the results obtained, during each examination during dynamic outpatient observation, the doctor makes the necessary additions and changes, determines treatment measures and the frequency of repeated examinations in accordance with the changes in the course of the disease, and also appoints the necessary consultations and additional studies (according to indications).
5. Conclusion
A DSS based on data mining has been developed to help the doctor in solving problems of planning and monitoring the timeliness, effectiveness and quality of dispensary monitoring of registered patients. Using a decision tree allows you to identify hidden relationships between parameters that affect the disease and the disease itself.

The use of IDM on data presented using OLAP analysis systems in the form of a real or virtual hypercube’s information is an effective tool for the effectiveness of the management for medical monitoring.

The results can be used for cloud technologies in educational institutions, approximation of the distribution law, switching subsystems within the framework of distributed operational annunciator and monitoring systems, SOA reference model, policy control at the network edge, air quality monitoring assessment based on IoT platform.

References
[1] Gusev A V, Romanov F A and Dudanov I P 2005 Review of medical information systems on the domestic market in 2005 Symp. “Medical information systems” (Kondopoga Publ.), 8-20
[2] Zhou L and Sordo M 2021 Expert systems in medicine Artificial Intelligence in Medicine. Technical Basis and Clinical Applications (Academic Press) chapter 5 pp 75-100
[3] Primova H, Sakiev T and Nabieva S 2019 Development of medical information systems Int. Conf. on Information Science and Communications Technologies (IEEE) pp 1-4 doi: 10.1109/ICISCT47635.2019.9011867
[4] Isabona J and Ojuh D O 2021 Intelligent Systems and Applications International Journal of Intelligent Systems and Applications 13(3) 35-42
[5] Lvovich Ya E, Tishukov B N, Preobrazhenskiy A P, Pitolin A V and Kravets O Ja 2019 Complex-Structured Objects Optimization During Modeling on the Population Algorithms Adaptation Basis *International Journal on Information Technologies and Security* 3 (11) 41-50

[6] Memeti A, Azizi A and Luma-Osmani S 2019 Human Resources Management System: SOA Reference Model *International Journal on Information Technologies & Security* 4 (11) 29-38

[7] Ivanova D and Elenkov A 2019 Big Data Analytics for Air Quality Monitoring Assessment Based on IoT Platform *International Journal on Information Technologies and Security* 2 (11) 43-50

[8] Romansky R 2019 A Survey of Informatization and Privacy in the Digital Age and Basic Principles of the New Regulation *International Journal on Information Technologies and Security* 1 (11) 95-106

[9] Patel A R and Patel J M 2012 Integrated approach of OLAP and Data Mining for Decision support *International Journal of Scientific Research* 2(3) 81-2

[10] Cheng D, Jian-Sheng Q, Ting Z and Wen-Jie L 2009 A high-speed response linkage model for integrated monitoring system at coal mine based on intellectualized data analysis *Procedia Earth and Planetary Science* 1(1) 1455-60

[11] He Z, Wu Z, Xu G, Zou Q 2021 Decision Tree for Sequences *IEEE Transactions on Knowledge and Data Engineering* PP(99) 1-1 DOI: 10.1109/TKDE.2021.3075023

[12] Bolnokin V E, Mutin D I, Mutina E I and Storozhev S V 2019 The synthesis of the algorithms for adaptive control by nonlinear dynamic objects on the basis of the neural network *IOP Conf. Ser.: Mater. Sci. Eng.* 537 042013

[13] Nedyalkov I, Stefanov A and Georgiev G 2019 Studying and Characterization of the Data Flows in an IP-Based Network *International Journal on Information Technologies and Security* 1 (11) 3-12

[14] Tsaregorodtsev A V, Kravets O Ja, Choporov O N and Zelenina A N 2018 Information Security Risk Estimation for Cloud Infrastructure *International Journal on Information Technologies and Security* 4(10) 67-76