Applying Data Mining Techniques When Making Medical Diagnostic Decisions*

Elena Mokina[1], Olga Marukhina[2], Mariya Shagarova[3]
[1,2,3] Dept. of Optimization Control Systems
[1,2,3]National Research Tomsk Polytechnic University
Tomsk, Russia

***Abstract*—Under the present-time conditions of the increased pace of life in large cities neurological disorders are tending to increase. The present paper considers the application of Data Mining techniques for studying medical data and building the decision support system on the basis of research results being, in the present case, the detection of the neurological disorders by the result indicators of the surveys on living standard, anxiety and depression. Throughout the use of Data Mining techniques there was built a decision tree and were established the reasoning rules, which provided the basis for the decision support system. The paper presents the basic requirements for this system enabling to reduce time of the clinical staff spent on processing survey data and providing recommendations on establishing diagnoses.**

Keywords—Data Mining, information systems, decision support system, SF-36, HADS_T.

I. INTRODUCTION

Nowadays, information is one of major world resources and information systems are an indispensable tool in practically all spheres, including medicine. Application programs and information systems capabilities in the sphere of medicine enable to modernize working process and improve the methods of diagnosing and curing. The relevance of the given paper is determined, on the one hand, by the necessity to develop a software product capable of supporting the process of diagnosing and analyzing health and living standard indicators and, on the other hand, the necessity to process and analyze the existing data arrays enabling to study patient data in dynamics while during the treatment indicators can have different values. To develop the decision support system it is necessary to establish the link between the diagnoses and indicators' values according to the assessment principles adopted in a certain healthcare organization.

II. PATIENTS' LIVING STANDARDS ANALYSIS

Research in this area is one of the major fields in modern medicine, which is confirmed by numerous surveys aimed at evaluating living standards, levels of patients' physical and psychological well-being. The surveys under discussion contain response options and are developed in the way to apply the method of ratings aggregation for making calculations. One of the best known surveys for studying patients' living standards is a short form of patient health status measure called SF-36, which came out of MOS (Medical Outcomes Study). SF-36 enables to register and quantify the changes in the living standards of patients with specific diseases during a definite period of inpatient treatment and to determine the components most favorable for the changes in their living standards caused by the treatment prescribed.

The choice of this survey by researchers (health care workers) is determined by the opportunities to use the obtained results for evaluating living standards of patients with any kind of disorders, to compare the outcomes with the statistics of the Russian Population Control concerning the corresponding groups as well as to evaluate the patients' living standard in a comprehensive manner (including social and psychological disorders).[1] In accordance with this study, to calculate the resulting indicator values (PH, a physical health component and MH, a mental health component) the values of the average deviation of population indicators are applied. It is stated that for different nosologies the resulting indicators of living standards are different as well. If while carrying out research a doctor does not know the average population values, it is recommended to use the earlier data gathered by the Russian Population Control concerning the corresponding groups. However, it is not always possible to provide accurate research results when using the available data because population data can vary in different geographic regions.

The integral estimation of living standards depends on a geographic region in which the research is carried out. Thus, to derive an estimate of the living standards of patients with certain types of nosology (in the present case, these are patients with neurological disorders) of a certain geographic region (in the present case, it is Tomsk region) it is practical to identify its population indicators [2]. Additional means such as Hospital anxiety and depression scale (HADS) aimed at assessing the reflection of tension symptoms are applied in the process of studying living standards. This questionnaire enables to assess anxiety as a passing state experienced in special situations (over the course of a disorder).

Diagnosis, including the one of neurological disorder, influences the indicators of living standards and patient's anxiety level. The identified dependences can be applied for establishing diagnosis in accordance with the conducted research on studying the indicators of living standards and the

The research is conducted with financial support from The Russian Foundation for Basic Research, project № 14-07-00675 and partially 14-06-00026

levels of anxiety and depression (basing on SF-36, HADS, etc.).

Support of establishing diagnosis in patients with a certain nosology via the decision support system includes such tasks as: to analyze patient's survey outcomes and compare them with the model of establishing diagnosis; to monitor the changes of a patient's living standard in dynamics, while additional diagnostics enables to improve treatment and foresee the changes of living standard indicators.

Calculation of living standards and health result indicators by means of different techniques (health surveys) is made with the accordance to the corresponding algorithms [3]. A short form of patient health status measure MOS SF-36 is one of the most labor-intensive ones and is aimed at getting the final calculations outcome. Thus, one more objective is to automate the process of handling data obtained from a patient who took the test and enabling to store the survey source data.

III. DATA MINING APPLICATION

The major objectives along the process of analyzing data for developing the intellectual component of decision support system are the following:

- to detect the hidden patterns of the existing data by means of Data Mining;
- to establish the rules of logical reasoning on the basis of Data Mining;
- to establish the rules of retrieving messages concerning the proposed diagnosis and the projected indicators change;
- to develop the logical rules of outputting the results of surveying in a programming language;
- to build a knowledge base.

After the processing of the accumulated data array it is possible to identify dependencies on their basis and to establish the rules of logical reasoning and to present data as the sentences of the following type: if (condition), than (action). One of the demonstrable ways of presenting such research results in intellectual data analysis is the decision trees whose interface is understandable both for an information technology specialist and a health care worker [4].

RapidMiner environment was used as a tool for a decision tree building and logical rules establishing. RapidMiner is a complex system which implements the methods of Data Mining (the methods of intellectual analysis) and statistical analysis, and owns a set of algorithms for processing and analyzing, including the ones for processing large data arrays. As well as with a constructor it is possible to operate with any set of data and add various operators of input and output, processing, visualizing, analyzing, etc. the whole process is presented in a tree form (Fig.1)

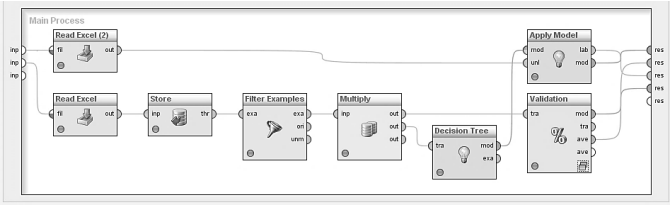

Fig. 1. Stages of analyzing process in Rapid Miner

The example of graphical representation of the anxiety detection decision tree (HADS_T) (depending on MH (a mental health component) and PH (a physical health component) indicators values), generated by Rapid Miner Miner in case of the absence of anxiety and depression, is presented both graphically and textually in Fig2 and Fig3.

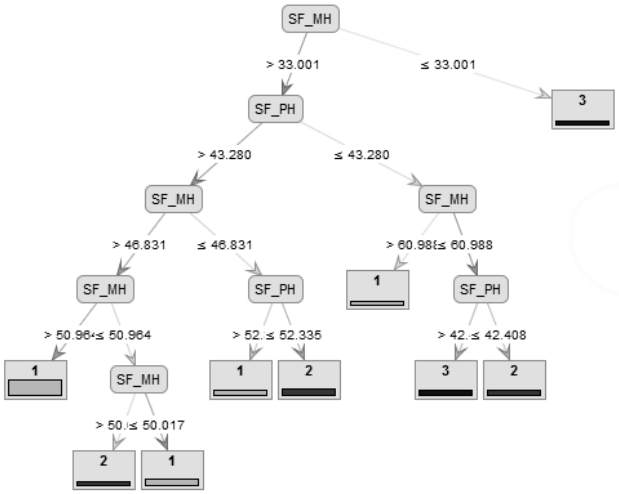

Fig. 2. Decision tree. Values 1, 2, 3 of HADS_T indicator are: 1 – normal; 2 – subclinically expressed anxiety/depression; 3 – clinically expressed anxiety/depression

```
SF_MH > 33.001
|  SF_PH > 43.280
|  |  SF_MH > 46.831
|  |  |  SF_MH > 50.964: 1 {3=0, 1=51, 2=0}
|  |  |  SF_MH ≤ 50.964
|  |  |  |  SF_MH > 50.017: 2 {3=0, 1=0, 2=2}
|  |  |  |  SF_MH ≤ 50.017: 1 {3=0, 1=11, 2=0}
|  |  SF_MH ≤ 46.831
|  |  |  SF_PH > 52.335: 1 {3=0, 1=6, 2=0}
|  |  |  SF_PH ≤ 52.335: 2 {3=0, 1=0, 2=11}
|  SF_PH ≤ 43.280
|  |  SF_MH > 60.988: 1 {3=0, 1=2, 2=0}
|  |  SF_MH ≤ 60.988
|  |  |  SF_PH > 42.408: 3 {3=6, 1=0, 2=0}
|  |  |  SF_PH ≤ 42.408: 2 {3=0, 1=0, 2=6}
SF_MH ≤ 33.001: 3 {3=5, 1=0, 2=0}
```

Fig. 2. Decision tree description

During the study of patients with neurological disorders and patients' living standards there was conducted the research on the dependency of the physical health component, mental health component, anxiety and depression indicators on the diagnosed case.

For an illustrative purpose, below is given a set of rules for diagnosis establishing according to physical health component (SF_PH), mental health component (SF_MH), the level of depression (HADS_D) and anxiety (HADS_T), diagnosis (D with the values "healthy", G20, G24, G35), where:

HADS_T > 14.500: G20 {G20=6, G35=0, G24=0, healthy =0}
HADS_T ≤ 14.500
| SF_PH > 43.231
| | SF_MH > 48.304
| | | SF_PH > 44.103
| | | | SF_PH > 60.168: G35 {G20=0, G35=2, G24=0, healthy =0}
| | | | SF_PH ≤ 60.168
| | | | | SF_MH > 53.077
| | | | | | SF_MH > 58.968: healthy {G20=0, G35=0, G24=0, healthy =9}
| | | | | | SF_MH ≤ 58.968
| | | | | | | SF_PH > 57.342: healthy {G20=0, G35=0, G24=0, healthy =4}
| | | | | | | SF_PH ≤ 57.342: G35 {G20=0, G35=17, G24=0, healthy =0}
| | | | | SF_MH ≤ 53.077: healthy {G20=0, G35=0, G24=0, healthy =24}
| | | SF_PH ≤ 44.103: G35 {G20=0, G35=5, G24=0, healthy =0}
| | SF_MH ≤ 48.304: G35 {G20=0, G35=25, G24=0, healthy =0}
| SF_PH ≤ 43.231
| | SF_MH > 60.988: healthy {G20=0, G35=0, G24=0, healthy =2}
| | SF_MH ≤ 60.988: G24 {G20=0, G35=0, G24=6, healthy =0}

IV. DECISION SUPPORT SYSTEM

In accordance with the stated tasks and requirements during the conduction of medical research it is necessary to develop the information support system maintaining the testing process by a user (a patient), result indicators calculation as well as establishing a diagnosis and projecting the indicators change. During this system development special attention should be paid to such functions as:

- a method of establishing diagnosis, determining the dynamics and characteristics of the indicators change on the basis of the survey conducted;
- processing the obtained data input by a user (a survey respondent) and presenting the results after each survey completion;
- storing the results obtained on the basis of the completed test in a systemized form;
- providing a researcher with the access to the results stored with the aim of data sampling in reliance on the required parameters;
- exporting the selected data, which a researcher needs for conducting analysis in other statistical programs.

When developing the information support system the following opportunities for ensuring system flexibility in case any changes could occur in specialists' work should be anticipated:

- changing/adding population values of indicators of a standard deviation and mathematical expectation;
- adding new surveys;
- adding new rules of diagnosis establishment output, projected change of the indicators for expanding the knowledge base;
- storing the number of the research conducted, its objective (in case of SF-36 the opportunity of taking the test should be presented in two forms: with the function of indicators calculation (PF, RP, BP, GH, VT, SF, RE, MH), without the result indicators (MH (a mental health component) and PH (a physical health component) calculation or with the function of diagnosis establishing).

The module for handling data obtained by surveying (carrying out research) must enable to make the calculation of:

- the resulting indicators values on a patient (opportunity to apply the necessary average population values);
- the average population values (with the choice of required records aggregate).

Proceeding from the requirements to the programming solution being developed, the model of the decision support system can be presented as follows (Fig.4):

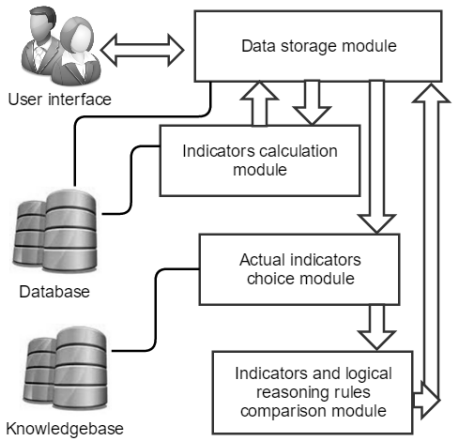

Fig.4. Decision support system

Subsequent to the questions answering, raw data (answers) are stored in the system and then the results processing is initiated: indicators are calculated by a definite algorithm (depending on the given answers) and are stored in a database [5]. In the following step (depending on which research in succession is being carried out) the process of actual values (not projected ones) is initiated. These data are compared with the rules of logical reasoning (a production rule) and the projected diagnoses as well as the indicators of the projected values are being output.

Knowledge base can be represented by a set of rules upon which the logic reasoning mechanism determines the output data. An approach based on the rules of logical reasoning was chosen as an approach to the development of the decision support system. The choice of this approach was made in view of the tasks stated (establishing diagnosis, analyzing data in dynamics), the required solutions, and the availability of the accumulated data on diagnoses, indicators referring to health condition of patients with disorders. Such rules enable to present knowledge as sentences of the following type: if (condition), than (action). After processing the accumulated data array it is possible to obtain the identified dependences and establish the reasoning rules on their basis.

An example of coding a rule is as follows:

IF HADS_T <= 14,500 AND SF_MH <= 60.988 AND SF_PM >42,408 THEN: D = G20

IF HADS_T <= 14,500 AND SF_MH <= 60.988 AND SF_PM <= 42,408 THEN: D = G24

Thus, it is possible to formulate all the gained rules whose body will become the knowledge base for the decision support system and programming the module “Comparing indicators with logical rules” enables to establish diagnosis, to predict indicators changes, and provides new reasoning.

In the process of projecting the database structure the following notions were taken into account:

- a respondent can pass surveys several times. A respondent can be subject to passing one and the same survey several times.
- surveys are made up of questions. Some questions belong to a definite category. Questions in the surveys are not repeated.
- one question consists of several response options (a user must choose only one, which is suitable for him or her). If several questions refer to one category, these questions can have the same responses.
- after the test is passed, the indicator is calculated on the basis of the responses.
- a question refers to the definite indicator. An indicator can consist of several questions. There is an indicator example: emotional state, moral state, etc.
- a value calculated for a respondent is within a definite numeric range.
- for each indicator there is a corresponding boundary defining whether this indicator is in the norm or higher/below the boundary of the normal value.
- one survey can be used for calculating several indicators.

V. CONCLUSION.

In accordance with the examined approach it is possible to conclude that the main process is the identification of the accumulated data dependencies and setting the logical reasoning rules, which are further applied in programming solution and data mining; and methods applied are crucial in this process. This is due to the fact that it is the accuracy of building a model and choosing criteria for analysis which the validity and value of recommendations reasoning for establishing diagnosis, monitoring health indicators in dynamics and system functioning depend on. Knowledge base is represented by production rules obtained by means of intellectual data analysis. The given paper suggests the approach to decision support system projecting with the aim to generate recommendations for diagnoses establishing on the ground of the obtained results.

REFERENCES

[1] Berestneva, O.G., Pekker J.S. Simulation and evaluation of biological systems adaptive capabilities // Proceedings of 2014 International Conference on Mechanical Engineering, Automation and Control Systems, MEACS 2014, 15 December 2014, Article number 6986860.

[2] Berestneva O.G., Marukhina O.V., Ivankina L.I., Shukharev S.O. Modelling of Adaptation Strategies for Different Entities // The European Proceedings of Social & Behavioural Sciences (EpSBS). — 2016. — Vol. 7 : Lifelong Wellbeing in the World (WELLSO 2015), — pp. 252-258.

[3] Mokina, E. Expert estimates in the informational support system of the university strategic plan// 8th Korea—Russia International Symposium on Science and Technology — Proceedings: KORUS 2004, — Vol. 3, — pp. 248—251.

[4] Marukhina O.V., Mokina E.E., Berestneva E.V. Using data mining for revealing hidden regularities in the task of analyzing medical data // Fundamental research. — Vol. 4-0—2015, —pp. 107-113.

[5] Meshcheryakov R.V., Balatskaya L.N., Choinzonov E.L. The Special Information System Supporting the Medical Institution Activities // Information - Control Systems. — 2012, — Vol. 5, — pp. 51-56.