Information system for monitoring of hypopituitarism in childhood and adolescence

M O Rusakov and O S Krotova

Department of Theoretical Cybernetics and Applied Mathematics, Altai State University, 61 Lenina ave., Barnaul, 656049, Russia

E-mail: mor1999@mail.ru, kr.olga0910@gmail.com

Abstract. The article is devoted to the development of a medical information system (MIS) designed to monitor and predict hypopituitarism in children and teenagers. MIS combines the medical information storage and management system with an intellectual system of medical decision-making support. The information system is based on a database with information obtained from impersonal medical discharge papers of children and adolescents of the Altai Territory diagnosed with hypopituitarism. Medical discharge papers provide information about the patient's health status while in the hospital, the results of examinations and treatment. Discharges form a complete medical history of the patient, and the information system allows assessing the disease in dynamics. The purpose of the intelligent information system is to predict the patient's growth in order to select the optimal treatment strategy. The intelligent decision support system is based on a machine-learning model trained on data from a database. The model was trained in the Python programming language. The MIS interface is developed using the C# programming language. The use of MIS in medical institutions will allow doctors to carry out a personalized approach to the monitoring and predicting the growth for patients, and to choose the optimal trajectory of treatment.

1. Introduction

Nowadays information technologies, artificial intelligence and big data analysis technologies are widely spread in different spheres of human activity. Health care refers to those areas of human activity in which a large amount of data must be analyzed to make decisions. The use of intelligent information systems in such areas can achieve a high economic and social effect. The volume of electronic medical data is growing every year, which provides great opportunities for the development and implementation of IT-solutions based on artificial intelligence. AI-powered software is able to combine large amounts of data with fast, iterative processing capabilities and intelligent algorithms, which allows programs to automatically learn based on patterns and features contained in the data. Artificial intelligence systems can be applied in all areas of medicine, including medical decision making, drug development, and patient care [1].

Hypopituitarism is a disease of the endocrine system associated with decreased or absent secretion of one or more pituitary hormones. Growth hormone deficiency is most often diagnosed in children and adolescents, which leads to slow overall development and stunted growth [2]. Growth hormone deficiency can occur both in isolation and in combination with a deficiency of other pituitary hormones [3, 4].
The aim of the study is to develop a medical information system for monitoring and predicting growth in hypopituitarism in childhood and adolescence. The Medical Information System (MIS) includes a system for storing and processing medical information and an intelligent medical decision support system. The main advantage of a complex MIS is to increase the effectiveness of the treatment of the disease.

The relevance and practical significance of the study is determined by:
- the need to achieve accelerated growth rates of the patient in the first years of treatment of hypopituitarism and their normalization in the future [5];
- the need to reduce the risks of complications;
- the need to use artificial intelligence technologies to automate decision-making processes in healthcare [6].

2. Medical Textual Data Processing and Database Development
The basis for the information system being developed is a database containing information about patients, medical examination results and treatment.

The initial data for the research are presented in the form of 500 medical discharge papers, containing impersonal information about patients and their hospital stay. A part of the data is presented in a table form, while others are medical records in natural language. Each discharge paper is filled in by the doctor in hand and, in case of hypopituitarism patients, contains the following sections: data about the patient, information on diagnosis and complications, disease and life anamnesis, examination results, received treatment, medical recommendations.

The first step to developing an MIS is to extract information from medical discharge papers and enter the data into a database [7]. Information extraction is one of the types of text analysis that allows you to extract relevant data from it and present it in a structured form. To extract information from semi-structured medical discharge papers, it is necessary to solve the following tasks:
1) extract numerical characteristics from tables;
2) extract numerical characteristics from the text;
3) find references to medical concepts in the text (diagnosis, concomitant diseases, complaints and symptoms, drugs, medical procedures), and extract fragments of text data.

Due to the fact that the templates for medical discharge papers have changed over the years, in order to avoid the loss of information when solving the assigned tasks, the following facts must be taken into account:
- discharge papers may contain various abbreviations;
- synonyms are used;
- tables containing the same indicators may have a different structure;
- the arrangement of indicators in the table is not in a certain order.

The high-level programming language Python was chosen to solve the set tasks. Python is one of the most popular programming language for data analysis and machine learning. The popularity of the language is partly due to the large number of modules and libraries that accelerate the solution of many applied problems of data science, computer vision and natural language processing. Information extraction is a natural language processing problem. Natural language processing is a direction of informatics at the junction of artificial intelligence and linguistics, associated with the study of methods of analysis and synthesis of natural language.

The python-docx library is intended for working with text files of the "\.docx" format in Python. The library provides many functions and methods for working with text files, including the ability to extract and process tables. Regular expressions, a formal language for searching and manipulating substrings in text, allow you to extract fragments from text. Regular expressions are implemented in the standard re module of the Python programming language.

Figure 1 shows a database model that reflects key entities - tables and relationships between them. The database consists of 17 tables and contains more than 150 attribute names. The database contains two types of tables: tables containing patient data that do not depend on the period of hospital stay (for
example, gender, age of the patient, etc.), and tables containing data related to the period of hospital stay (results analyses and examinations, medical treatment, etc.). The patient ID is used to link the tables of the first type, and the statement ID is used for the tables of the second type.

Figure 1. Database ER-diagram.

3. MIS Functional Characteristics
The purpose of the information system is to provide endocrinologists with efficient tools for effective monitoring of the disease course.

- MIS will provide a quick approach to hospital discharge papers, formal medical records to fix patient's physical growth values on percentile curves;
- MIS will make it possible to predict patient's growth by means of the intellectual decision-making support system based on machine learning models. Due to the intellectual system, a doctor will be able to choose the optimal treatment strategy according to the patient's predicted height.

On June, 17 the Ministry of Justice of Russia registered the order of the Ministry of Health No. 911n "On the approval of requirements for state information systems in the health care field of the subjects of the Russian Federation, medical information systems of medical institutions and information systems of pharmaceutical institutions". The order determines a mandatory set of functions for information systems to provide medical institutions, among them conducting e-health records; centralized systems (subsystems) of storage and processing the results of diagnostic and laboratory tests, implementation of telemedical consultations, etc. In accordance with the order, MIS is focused on collective use by endocrinologists of one medical institution and can be integrated into the information system of a medical institution as a separate module intended for narrow specialists.

The information system has a module structure where each module is designed to solve a definite set of problems. MIS includes the following components:

1. Data Storage is a component intended for storage structured medical data.
2. The "Patients Database" module allows keeping a single registry of patients; viewing, editing, and adding information on patients.
3. The "Health Records" module is used for viewing, editing, adding and deleting information on a patient's hospitalization period (Figure 2).
4. The "Patient's Physical Growth Rates Control" module enables doctors to mark the values of a patient's physical growth on percentile curves, as well as to monitor changes in dynamics. The percentile curves for controlling the growth of boys and girls are different. Figure 3 shows information system window for controlling the growth of boys.
5. The "Disease Prognosis" module is a decision-making support system designed for predicting the patient's growth by means of machine learning models as well as for selecting the optimal treatment for the patient.

6. The information security subsystem is aimed at providing information security by means of permissions access to the information system data.

The object-oriented programming language C# was chosen for the implementation of the software interface.

Figure 2. «Health Record» module interface. This information system window allows doctors to view, edit, add and delete information about the patient's hospitalization.

4. Medical decision-making support system

The intellectual decision-making support system integrated into MIS is designed for predicting patient's growth in order to choose the optimal treatment strategy. The basis of the intellectual decision-making support system is the model received as a result of machine algorithm learning on MIS data. The appliance of machine learning techniques in medicine is a method of a personalized approach to diseases diagnosing and prognosis. Personalized medicine is a new organizational model of health care which is based on selecting individual medical, diagnostic, and preventive means best for the patient in order to exclude negative results and complications [8].

The objective of predicting growth for children and teenagers with hypopituitarism is that of regression modelling. In a classical regression problem, the training set is presented by a number of separate objects characterized by a feature vector. Continuous real variable $t$ serves as a dependent (predicted, target) variable. The task is to build an algorithm which will return the point estimate of regression value by instance description $x$.

The dependent variable in the problem under consideration is the patient's growth at the next hospitalization after a treatment course (5-7 months) of "Растан®" (Rastan®). The set of independent (input) variables is determined in the course of the study by means of correlation analysis.

Correlation analysis is closely connected to regression analysis. With the help of the latter, one can determine the necessity of entering certain features in the regression equation. The more the correlation coefficient modulo between the feature and the target variable is, the more informative the feature is. Independent values correspond to a zero correlation coefficient, while the extreme values of the correlation coefficient (+1 and -1) correspond to the existence of linear relationship which is the strongest of all possible relationship forms.
Figure 3. A window for capturing patient's physical growth values on percentile curves. Assessment of the level of physical growth according to percentile curves of height and weight is assessed by comparing the age (lower scale) and the height or weight of the child (side scale).

Based upon correlation analysis we selected the most informative features: patient's current height, weight, age, level of Insulin-like Growth Factor 1 (IGF-1) and the level of creatinine in the blood. Table 1 demonstrates the correlation coefficients between the dependent variable and the selected independent variables.

Table 1. Correlation coefficients values of the selected independent variables with the target variable.

| Feature                   | Correlation coefficient |
|---------------------------|-------------------------|
| Patient's current height  | 0.916                   |
| Patient's weight          | 0.829                   |
| Patient's age             | 0.677                   |
| IGF-1                     | 0.696                   |
| "Rastan®" dosage (mg)     | 0.805                   |
| Blood creatinine level    | 0.562                   |

For designing the intellectual decision-making support system aimed at patient's growth predicting it was decided to learn and evaluate the quality of several models: multi-variate linear regression, support vector machine, and decision trees. Programming language Python and Scikit-learn library were chosen for modelling. Scikit-learn is the most reliable library for machine learning in Python. Scikit-learn contains a great amount of modern supervised and unsupervised machine learning algorithms as well as full documentation for each of them [9].

Multi-variate linear regression. A multi-variate linear regression is the one in the model of which objects and features are n-dimensional vectors. A multi-variate linear regression differs from an univariate one by the hyperplane used instead of a regression line. For evaluation of regression line parameters the method of least squares is applied. Regression line modelling in Python can be performed by means of LinearRegression() class of the linear_model module in Scikit-learn library.
SVM regression. Support vector machine (SVM) is a universal set of machine learning algorithms which are used for solution of linear and non-linear problems of classification and regression.

The support vector machine solution of linear classification problems lies in building the decision boundary so that the training set objects would be located at the greatest distance from it [10]. To provide this two parallel hyperplanes are built on the both sides of the decision boundary. The hyperplanes are "supported" by the nearest objects of each class. The algorithm requires an assumption that the longer is the distance between the planes the less will be the average classification error.

In the regression problem, the algorithm attempts to place as many objects between the hyperplanes as possible. The Scikit-learn library of Python language contains a number of classes by to build the following support vector machine: LinearSVR(), SVR() and NuSVR(). LinearSVR() was chosen to solve the problem of predicting the growth. To solve the problem of forecasting growth, the LinearSVR() model was chosen.

Decision Trees. Decision trees make it possible to solve problems of classification and regression by means of logical frameworks. Based on features of the training set data the decision tree model learns based on the hierarchically organized question system. Meanwhile, the question asked at each following hierarchical level depends on the answer received at the previous level. Decision trees are applied for solving various practical problems since they have a number of advantages: can work with data of any type, do not require data pre-processing, are easy to interpret, and allow selecting automatically the most relevant model features. Decision trees for the problems of classification and regression are realized in DecisionTreeRegressor() class of the tree model in Scikit-learn library.

To optimize the hyperparameters of the models, a grid search was used, which is implemented in the GridSearchCV() class of the sklearn.model_selection module.

For the quality assessment of the trained models, the coefficient of determination (R^2) was chosen. In assessing regression models the value of the determination coefficient is interpreted as a fit of the model to the data. The determination coefficient can vary from 0 to 1. For acceptable models the determination coefficient is meant to equal to not less than 0.5 (50%). The models with the determination coefficient of more than 0.8 (80%) can be estimated as good enough.

The determination coefficient values for each model are given in table 2.

| Model                           | R^2  |
|--------------------------------|------|
| Linear Regression              | 0.86 |
| Linear regression SVM          | 0.93 |
| Regression decision trees      | 0.85 |

High values of the determination coefficient point to the good fit of the models to the data. SVM has the highest value of the determination coefficient (93%). The model became the basis for a medical decision-making support system that is integrated into the MIS.

5. Conclusions
The increasing volume of medical e-records provides the opportunity to analyze the course of various diseases, to apply modern technologies and approaches of data analysis and artificial intelligence for studying medical decision-making processes. Application of artificial intelligence and implementation of medical decision-making support systems in medical institutions will provide the improvement of disease diagnosing in challenging diagnostic cases, selection of the optimal treatment strategies and therapy results forecasting [11].

In the course of the research, the goal was achieved and all the tasks were solved:
– text medical discharge papers were processed, information was extracted;
– a database has been created;
MIS has been developed, which includes a data storage system and a medical decision-making support system that will allow endocrinologists to effectively monitor and predict the growth of children and adolescents with hypopituitarism, and choose the optimal treatment strategy.

References

[1] Subramanian M, Wojtusciszyn A, Favre L, Boughorbel S, Shan J, Letaief K.B, Pitteloud N, Chouchane L 2020 *Journal of Translational Medicine* 18 (1) 472

[2] Yeliosof O, Gangat M 2019 *Current Opinion in Pediatrics* 31 531–536

[3] Ibáñez L, Barouti K, Markantes G, Armeni A, Georgopoulos N 2018 *Hormones (Athens)* 17 439–449

[4] Richmond E, Rogol A D 2016 *Best Practice & Research Clinical Endocrinology & Metabolism* 6 749–755

[5] Nagaeva E V 2013 *Problems of endocrinology* 6 27–43 (in Russian)

[6] Gusev A V, Zarubina T V 2017 *Doctor and information technologies* 2 60–72 (in Russian)

[7] Moskalev I V, Krotova O S, Khvorova L A, Bobkova D G 2020 *Journal of Physics: Conference Series* 1615 012031

[8] Johnson K B, Wei W Q, Weeraratne D, Frisse M E, Misulis K, Rhee K, Zhao J, Snowdon J L 2021 *Clinical and Translational Science* 1 86–93

[9] https://scikit-learn.org/stable/

[10] Battineni G, Chintalapudi N Amenta F 2019 *Informatics in Medicine Unlocked* 16 100200

[11] Hong N, Park H, Rhee Y 2020 *Endocrinology and Metabolism* 35 71–84