Application of data mining methods in solving medical problems

D K Muhamediyeva, O B Ruzibaev and Yu Baxromova

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Amir Temur street, 108, 100200, Tashkent, Uzbekistan
matematichka@inbox.ru

Abstract. Mathematical methods in medicine - the criteria for determining the status of research objects related to medicine and health problems, methods and algorithms for assessing, analyzing and forming a set of quantitative variables are analyzed. Because of the analysis, problems and tasks were identified that need to be addressed.

1. Introduction

Mathematical methods in medicine are a set of methods and algorithms for determining the status of research objects related to medicine and health problems, assessing, analyzing and forming a set of quantitative variables [1].

Mathematical representation of processes, the use of mathematical methods is carried out mainly according to two directions. The first is a statistical method of processing medical data, and the second is the construction of mathematical models of the process of determining, assessing, diagnosing and classifying the state of the analyzed medical objects.

Medical processes, especially the state of normal and pathological activities of the body, the processes of diagnosis and treatment, as well as the body systems as a whole, are easy to represent using mathematical methods [3].

Ensuring the reliability of the properties of a medical object, practical conclusions, recommendations, methods of disseminating the studied data in a series of observations and using the appropriate mathematical apparatus is carried out on the basis of certain hypotheses. The definition of the description of the properties of an object, determined using mathematical statistical methods, are called data generation models. Data models do not contain any information or hypotheses about the internal structure of a real object and rely only on the results of experimental, instrumental measurements [2].

In both cases, the direction depends on the system models based on the mathematical representation of meaningful information about the structure of the system under study objects and events as well as the mechanisms of interdependence of individual elements. The development and practical application of mathematical models of systems (mathematical modeling) is one of the promising areas of using mathematical methods in biology and medicine [3,4].

The modeling process begins with the construction of simplified cases and their mathematical expression, which, with one hand, it reflects a qualitative indicator of the main events, on the other hand, it allows a very simple mathematical expression of processes. As the research deepens further, new models will be built for a detailed description of reality, event, process. At this stage, the secondary listed factors are removed [5].
Physicians typically conduct curriculum-based research using the mathematical apparatus of statistical theory to make it possible to use sample research to transfer data from research objects to general collections. In this case, the doctor must not only use the mathematical formula, but also draw conclusions from the methods for assessing the reliability of the data obtained on its basis. There are parametric and nonparametric types of methods for assessing the reliability of data [5, 6]. Methods of statistical data processing that require knowledge of the distribution law and the calculation of their main parameters are called parametric. Additional Student's t-tests are used to select related and unrelated data [2-4].

On the other hand, nonparametric methods do not require knowledge of the distribution law when forming the studied set of features and when selecting information. Nonparametric methods include: Mann-Whitney U-test (two unrelated selection methods), Wilconson test (for related data groups), Pearson compatibility, etc. [3-6].

Initial list of references studied in modern research, includes all clinical and laboratory indications that a researcher can use. In the volume of such data, first of all, the development of methods for processing multiparameter data began [6-11].

2. Description of medical dynamic processes using differential equations

Differential equations are widely used to describe dynamic processes and variables. For example, a linear differential equation called the elastic reservoir model describes the course of processes in the cardiovascular system as follows [7-9]:

\[
\frac{1}{k} \frac{dy}{dt} = \frac{y}{R} + W(t),
\]

where, \( y \) - instantaneous blood pressure value; Parameters R and k represent the total resistance of the vessel to arterial pressure and the coefficient of elasticity of the aorta; \( W(t) \) - instantaneous volumetric velocity of blood flow outgoing from the heart.

Therefore, below we will consider the Cauchy problem of the form

\[
\frac{dy}{dt} = f(t, y), \quad 0 \leq t \leq T,
\]

\[
y(0) = y^0.
\]

Main feature of problem (1) is that it is considered on a fixed time interval - it is included in the formulation of the problem. This formulation is natural for applications where problem (1) can act as a mathematical model. The classical existence and uniqueness theorem [7], which is local and guarantees the existence of a solution in some sufficiently small neighborhood of the initial point, as a rule, becomes of little use.

Let \( \varphi_1(t) \) – first negative root, \( \varphi_2(t) \) – first positive root, and let \( \varphi_1 = \max_{[0, T]} \varphi_1(t), \ \varphi_2 = \min_{[0, T]} \varphi_2(t) \).

Let also the initial value \( y^0 \) satisfies the condition \( \varphi_1 < y^0 < \varphi_2 \), and therefore,

\[
\exists \epsilon > 0: \varphi_1 + \epsilon < y^0 < \varphi_2 - \epsilon.
\]

Let's choose the lower solution \( \alpha = \varphi_1 + \epsilon \), and the upper solution \( \beta = \varphi_2 - \epsilon \). Because \( f(\alpha) > 0 \), a \( f(\beta) < 0 \), corresponding differential inequalities are satisfied. It follows from Chaplygin’s theorem that there exists a solution to the problem under consideration \( y(t) \), satisfying the inequalities \( \varphi_1 < y(t) < \varphi_2 \) [7-8].

Another example of the application of mathematical methods in medicine is the solution to the problem of the spread of an epidemic. Over the many millennia of human existence, a huge number of people have died from various epidemics. In order to be able to fight epidemics, that is, to timely apply certain medical measures (quarantines, vaccinations, etc.), it is necessary to be able to compare
the effectiveness of these measures. They can be compared only if it is possible to predict how the course of the epidemic will change, especially the number of patients, in a given event. Hence, it becomes necessary to build models that could serve the purposes of the forecast.

First, consider a model of the “natural” course of the epidemic (without medical intervention). It is clear that the epidemic model can include the influence of factors of various levels. So, one could take into account the laws governing the activity of bacterial cells, the degree of susceptibility to infection of individual people, the likelihood of infection carriers meeting with still healthy people, and many other factors. Since our goal is only to create an illustrative model, we abstract from many factors.

Let there be \( N \) healthy people, and at time \( t = 0 \) one sick person (source of infection) falls into this group. Suppose that no removal of the sick from the group occurs (there is no recovery, no death, no isolation). We will also assume that a person becomes a source of infection immediately after he himself becomes infected [9-12].

Let us denote the number of sick people at time \( t \) by \( x(t) \), and the number of healthy people by \( y(t) \) (obviously, \( x(t) + y(t) = N+1 \) at any time).

At \( t = 0 \), the condition \( x(0) = 1 \) is satisfied.

Consider the time interval \( t + dt \), where \( dt \) is a small time interval. It is necessary to determine how many new patients will appear during this period of time. It can be assumed that their number will be proportional to the value of \( dt \), as well as the number of meetings of healthy and sick people, that is, the product of the values \( x \cdot y \):

\[
dx = \alpha xy dt,
\]

where \( \alpha \) is the proportionality coefficient (transmission coefficient).

\[
y = N + 1 - x \rightarrow \frac{dx}{dt} = \alpha x(N + 1 - x).
\]

Solution to this equation:

\[
x(t) = \frac{N + 1}{Ne^{-\alpha(N+1)t} + 1}.
\]

You can complicate the model by assuming, for example, that at time \( t \), not one person is sick, but several \((b)\). In addition, suppose that after a short period of time the patient recovers and receives immunity. Then \( z(t) \) is the number of recovered and vaccinated patients who have recovered and by time \( t \).

\[
x + y + z = N + b,
\]

\[
\begin{cases}
\frac{dx}{dt} = \alpha xy - \gamma x, \\
\frac{dy}{dt} = -\alpha xy,
\end{cases}
\]

where \( \gamma x \) is the number of recovered.

The model can take into account mortality from disease, transmission of disease through a vector (rodents), etc.

The spread of coronavirus infection in Uzbekistan may slow down by the end of summer, when many residents of the country will develop sufficient immunity.

All are either ill or vaccinated. Herd immunity can be calculated as follows: the number of detected infected multiplied by 10. This will be the true number of those who have recovered. To this must be added the vaccinated and divided by the number of inhabitants. And we get a figure of 70% of the population.

According to the immunologist, there will no longer be any rise in the incidence in the country.

There can be a rise in the only case: if the virus acquires a mutation that will give it a completely new function and it will be a new virus that will not be affected by the immunity gained during the pandemic.
3. Description of medical processes using statistical methods

Data mining is used in areas related to statistical techniques, hidden pattern detection, and artificial intelligence techniques in forecasting. The creation of automated systems is important when identifying logos. The creation of clear and effective classifiers for a large dataset is one of the key challenges of scientific research in the field of data mining. Currently, various classification methods have been developed, including decision trees, Bayesian method, logical regression method, SVM, KNN, etc.

One of the main principles of factor analysis and multivariate scaling (base component method) is to reduce the sample size based on creating a linear combination of all analyzed variables (the number of variables is set by the researcher and can be reduced to 2). [7]. In this case, the task is to preserve the maximum amount of information included in the entire data set, and an additional characteristic is included - the explained level of variance. The resulting factors include information that is present in the original variables, so that the number of variables is reduced (for example, to two). Typically, factor analysis attempts to provide a qualitative description based on a set of many parameters that have a goal. That is, to "generalize" information and obtain such factors, they must have a biological meaning. An interesting aspect of factor analysis is that it is "neutral" in terms of object analysis, since in practice the basic analysis is performed on a correlation matrix that does not include information about each object. This is important if the researcher wants to use the eigenvalues of factors in classification methods for further analysis. The use of such approaches helps to find solutions to modern problems with dimensional problems [2-6].

The famous Bayes theorem can be used to calculate the posterior probability of examination results for the presence of certain symptoms in the diagnosis of a patient's disease, for which the following scheme is used [1-3]:

The patient may have $S_1, S_2, \ldots, S_m$ illness.
- Each disease is given with a priori probability $P(S_i), P(S_2), \ldots, P(S_m) = 1$.
- Each disease is characterized by $C_1, C_2, \ldots, C_n$ symptoms, while each disease is characterized by "its own" $P(C_j / S_i), i = 1,2,\ldots, m; j = 1,2,\ldots, n$ symptoms in the "general" list.
- There is a conditional probability of the onset of symptoms for each disease.
- In this case, the prior probability of the presence of a disease for these symptoms can be identified using the following formula:

$$P(S_i / C_j) = \frac{P(C_j / S_i)P(S_i)}{\sum_{i=1}^nP(C_j / S_i)P(S_i)}$$

4. Computational experiment

The software was developed on the basis of scientific and practical results obtained in the framework of research work. By far the most widely used model problems are issues such as the diagnosis of diabetes in women (Pima Indians Diabetes) and the Haberman's Survival Data Set. Preliminary estimates for these issues, the lookup table data will be available on the Internet at http://www.ics.uci.edu. The Breast Cancer Database was created using data from William H., a physician at the University of Wisconsin at Madison Hospital. [1-9] have been used to describe symptom objects. Each element belongs to one of two possible classes: unsafe or unsafe. Number of subjects - 9, number of symptoms - 699. Distribution of classes - 458 (65%) and 241 (34.5%).

The problem of diabetes (Pima Indians Diabetes). In this regard, the process of diagnosing diabetes in women is discussed. The results of medical examinations of 768 women over the age of 21 are presented. The follow-up process was carried out by the American Institute of Diabetes and Kidney Disease in accordance with the requirements of the World Health Organization. In this case, 8-digit
values are given, which may indicate diabetic disease, and based on these values, the patient is
diagnosed with the presence or absence of diabetes.

   Signs important for the diagnosis of diabetes include:
   To solve the problem, you can get the values of one parameter - diabetes - 1 or diabetes - 0.
   We are required to build a fuzzy treatment and diagnostic system that gives the least number of
   errors. The training will be used to train a system of 450 objects in the sample, and the rest from the
   testing process.
   Haberman’s Survival Data Set. The Habermann problem arose from a study conducted at the
   Billings Clinic of the University of Chicago between 1958 and 1970. It addresses the question of
determining the survival rate of a patient with breast cancer after surgery. In the database there are 306
cases used in the case, each of which is characterized by the following characters:

   1. age (numerically) of the patient at the time of surgery;
   2. years of work of the patient (numerically);
   3. muscles out of shape (numbered);
   4. status of residence (class mark):
      1 - the patient lives for 5 years or more;
      2 - the patient dies within 5 years.
   As you can see, in the Haberman problem, objects are divided into two class. Training is used to
   train a system of 200 objects in the sample, and the rest of the objects are used for testing.
   We use 180 objects in the training set to train the system, and the rest of the objects are used to test
   the system. Curriculum selection data on these issues are tabulated in the appendix.
   Selection data for these samples is also shown in Table 1.

   Table 1. Formal presentation of test problems

   | Dataset name | Number of classes | Number of features | Number of objects |
   |--------------|-------------------|-------------------|------------------|
   | Haberman     | 2                 | 4                 | 306              |
   | Breast cancer| 2                 | 9                 | 699              |
   | Diabetes     | 2                 | 8                 | 768              |

   To solve the problems, a number of existing algorithms were used in order to conduct a
   comparative analysis of the results obtained. Table 2 below shows the results obtained using the
   proposed Apollo sphere decision rule method and the results obtained using the classification
   algorithms. The results of the comparative analysis showed that the methods and algorithms developed
   for solving the test problems of Haberman, Diabetes and breast cancer were more effective
   (highlighted in dark color) than the results of a number of existing classification algorithms (SVM,
   KNN, Naive Bayes). Results are based on software developed in the Visual C# programming
   environment.

   Table 2. Comparative analysis of the results of the decision rule method with existing
   algorithms

   | Name of the task | “Apollo ball” | SVM  | KNN  | Naive Bayes |
   |------------------|---------------|------|------|-------------|
   | Diabetes         | 93.21         | 92.33| 91.12| 89.62       |
   | Haberman         | 93.5          | 92.11| 90.7 | 88.56       |
   | Breast cancer    | 94.72         | 91.5 | 93.5 | 92.89       |
5. Conclusion

Improvement of methods and algorithms for the formation of diagnostic symptom complexes made it possible to effectively solve the problems of the subject area on the classification and clustering of objects. A numerical method of partial selection and an algorithm for choosing a set of informative signs corresponding to the classes of choosing a curriculum using the decision rule for classifying informative signs have been developed. The result made it possible to classify complexes of symptoms depending on the degree of the disease in the early detection and diagnosis of medical diseases and in making the right decisions regarding treatment.

A hybrid algorithm for the classification of informative symbols created using the KNN and Bayesian methods for identifying symbols has been developed. As a result, the methods of detecting the emblem have made it possible to increase the accuracy of early diagnosis, reduce the time of diagnosis using computer diagnostics, and treat the disease at an early stage. As a result, the rate of early detection of the disease increased and the time of diagnosis using differential computer diagnostics was reduced.

References

[1] Duke VA 2003 Information technology in biomedical research (SPb. Peter)
[2] Bates D W, Cohen M and Leape L L 2001 Reducing the frequency of errors in medicine using information technology ed J Amer (Med. Infor. Assoc) vol 8 pp 299–308
[3] Armitage P 2001 Theory and practice in medical statistics Statistics in Medicine vol 20 Issue 17-18 pp 2537–2548
[4] Zagoruiko NG 1979 Empirical prediction (Novosibirsk Science) p 120
[5] Dyukov E V and Zhuravlev Yu I 2000 Discrete Analysis of Feature Descriptions in High-Dimensional Recognition Problems Zhurn.Vych.Matem. and Matem. Physvol 40 No 8 pp 1264–1278
[6] Dyukov E V and Peskov N V 2001 On discrete recognition procedures based on the construction of class coverings (MMPO-X Moscow) pp 48–51
[7] Samarskiy A A 2002 Math modeling (FIZMATLIT)
[8] Aripov M 1988 Method of reference equations for solving nonlinear boundary value problems (Tashkent, Fan) p 137
[9] Belotelov N V and Lobanov A I 1997 Population models with nonlinear diffusion Math modeling No 12 pp 43–56
[10] Muxamediyeva D K 2019 IOP Conf.Journal of Physics: Conf. Series 1210
[11] Juraev Z Sh, Muhamediyeva D T and Sotvoldiev D M 2020 IOP Conf. Journal of Physics: Conf. Series 1546(1) 012083
[12] Mukhamedieva D T 2020 Journal of Physics: Conference Series1546(1) 012091