RESEARCH OF MEDICAL DIAGNOSTIC DATA SEARCH METHODS

Oleksii Vasilenko, Oleksandr Kuzomin, Oleksandr Shapoval

Abstract: *Improving the efficiency of diagnosis and treatment is the most important task of health care. One of the solutions to this problem is the complex automation of the collecting, storing and processing of medical information, for which medical information systems are being created. As well as advisory assistance to the doctor for the diagnosis of diseases, for which medical expert systems are created. The aim of the work is to research the methods of modeling and creating medical expert systems, the analysis of their disadvantages and the way of their solutions. Targeted research relies on the Bayesian trust networks.*

Key words*: medical expert systems, diagnosis, medical knowledge base, syndromes, sym*

Introduction

The general problem of identifying the diagnosis of a person's illness primarily and foremost lays on the presence of uncertainty, inaccuracy and insufficient volume of biomedical data and knowledge, difficulties in formalizing knowledge, excessive variety of making decisions regarding diagnosis, modeling a patient's condition, modeling the clinical thinking of a physician, choosing treatment methods, etc.

The object of the research is the processes of modeling the diagnosis of clinical medicine, the development of methods for the analysis and synthesis of biomedical data and the identification of knowledge from them and the creation of medical expert systems (MES) of clinical medicine.

The subject of the research is the models and methods of analysis of medical data and knowledge, the reliability of storing and processing of complex biomedical Big Data and the development of medical expert systems for the diagnosis of clinical medicine.

In medicine, it is important to find accurate methods for describing the data for research, as well as estimating and monitoring the process of diagnosis. The best way to the accuracy and logical considerations in solving any problem is to use a mathematical approach. This approach is might be chosen regardless of how difficult and complex the issue is. If we deal with a large number of interdependent factors, symptoms, syndromes, signs of illnesses that exhibit significant natural variability, then there is only the one effective way to describe the complex scheme of their effects – to the use of the appropriate statistical method. If the number of factors or the number of data categories is very large, then it is desirable or even necessary to use data processing methods such as Data Mining, in order to obtain the desired results in a short time, it is necessary to create and use medical expert systems (MES).

Research of the subject area and the choice of methods and models of medical diagnosis

It is common knowledge that the medical expert system is a set of programs that performs the functions of an expert in solving medical diagnosis problems. Perform an analysis of some well-known modern MES:

1. The system of medical diagnostic Diagnos.ru.
2. Diagnostic decisions of the expert system "ЕСБАД".
3. Expert system "МУТАНТ" (MUTANT), which was created by the staff of the Electronic Computing Center of Moscow University.
4. Automated system of early diagnosis of hereditary diseases "ДИАГЕН" (DIAGEN), which allows identifying more than 1200 forms.
5. Dendral - analysis of mass spectrometry data.
6. Mycin - diagnosis of infectious diseases of the blood and the antibiotics recommendations.
7. STD Wizard - an expert system for recommending and selecting medical analyzes (diagnostics).

As you can see in the analysis, it is possible to make three main conclusions:

1. Scope of application of each MES is objectively due to a narrow list of diseases, that is the representation of databases and knowledge of diseases.
2. On average, the MES have from 56% to 90% of the correct diagnoses.
3. The main principles for the development of the MES are production rules, which give a very large percentage of correct diagnoses.

These conclusions provide a basis for use in the development of MES to consider the possibility of using production rules in the creation of the MKB (medical knowledge base) for the development of MES. In addition, the data (Tab. 1) determine the most effective directions:

1. Using electronic medical data (EMD).
2. Building a knowledge base (KB) on the rules of products of the type "IF - THEN".

The analysis of the using of modern researches in the MKB development that were presented above provides a basis for the selection, modification or combination of methods for analyzing medical data for the diagnosis of diseases. When developing models and methods of diagnosis, one should take into account rather high results in the development of modification of statistical analysis methods and the application of Data Mining methods, which are often been supplemented by the use of fuzzy, hybrid neural networks in the modeling of complex applications. In addition, special attention must be paid to the very efficient use of belief networks that use the theory of subjective probabilities, are based on the Bayesian method, the Wald method, or, as it is still called, the method of sequential statistical analysis, the diagnostic tables of Sano, the method of linear discriminant functions, etc.

The use of subjective expectations in Bayesian networks is the only alternative in practice, if it is necessary to take into account the views of experts (e.g. physician) about the possibility of an event occurring to which the notion of repeatability is used and it is impossible to describe it in terms of a set of elementary events.

With these methods, one can calculate the distribution of probabilities in expert systems, which is a more convenient method for calculation, rather than assume with the help of statistical functions. With Bayesian theory, one can calculate the probability of judgments that are not certain. Such probability is determined by the level of confidence in the truth of a judgment. However, no matter how good this method seems, there are some negative aspects in using this theory. For example, in many cases it is psychologically difficult for an expert to remain within the framework of a strict mathematical apparatus of probability theory, which in its nature is objective. It is necessary to break the strict conditions of equality of units of probability sum of all possible states, especially in their large numbers. In many cases, actually observing evidence is confirmed not by any particular result (or hypothesis), but immediately by a certain set that does not allow determining the probability of each of them. If the expert estimates values that have a rather obscure meaning, whose properties in many cases do not coincide with the usual representations, it can being confronted with the fact that its answers will not provide useful informing about the estimated values.

Expert systems that use the theory of subjective probabilities are in great demand among expert systems for finding solutions. These expert systems allow you to get a right answer to various questions in a narrow subject area.

The most effective areas of data analysis and knowledge discovery in the proposed study, that was determined in (Tab. 1), are:

1. Method of using "micro situations" and theory of utility.
2. Fuzzy and hybrid neural networks.
3. The use of ontology, case studies and disease knowledge.

To develop models and methods of diagnosis, we need to make some clarifications regarding what clinical thinking is. Having determined it, we can more accurately develop a diagnostic algorithm for diseases. In addition, the basic logical connections of concepts reflecting the subject area and necessary to refine the development of models for clinical diagnosis should be used.

Clinical thinking in accordance with is the process of thinking physician from the moment of meeting with the patient or receiving the first preliminary information until the recovery or death of the patient. The result of the clinical thinking of a doctor is the formulation of clinical diagnosis, treatment plan and its practical implementation. Therefore, when we have plenty of previous examples of diagnoses (in ontological or precedent presented in the database and knowledge base) as the results of practical and positive examples of clinical thinking of doctors, we will be able to develop algorithms for automatic diagnosis of the disease.

The general version of the stages of creating a criminal diagnosis based on clinical thinking is depicted on.

With the development of medical science, a scheme for formatting a clinical diagnosis was developed:

1. The main disease.
2. Complications of the main disease.
3. Concomitant diseases.

In addition, it is advisable to take into account the following stages of the diagnosis (Fig. 1):

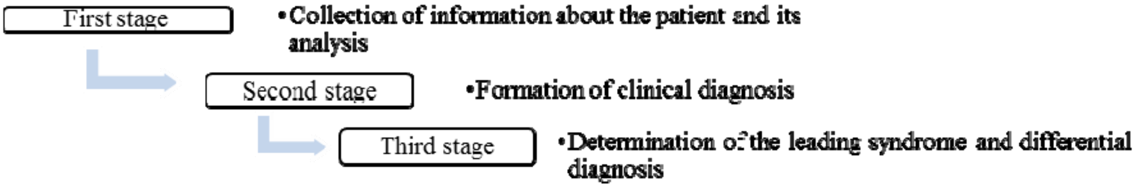

First stage
•Collection of information about the patient and its analysis
Second stage
•Formation of clinical diagnosis
Third stage
•Determination of the leading syndrome and differential diagnosis

Figure 1 – Diagnosis search algorithm

Table 1 – An example of the forming of clinical analysis

The main disease	Measles, typical, medium-grained form, period of rashes, abrupt course of the disease.
Complications of the main disease	Laryngitis
Concomitant diseases	Atypical dermatitis, localized form, remission period

As a variant of the stage of forming a clinical diagnosis (Fig. 1), it is expedient to bring the following:

1. Nosological form.
2. Degree of gravity of the process.
3. Stage, period of illness.
4. Degree of compensation.
5. Phase of the process (active, inactive).
6. The nature and localization of the pathological process.
7. Etiology of the disease.

However, these clarifications do not affect the concrete formation of the clinical diagnosis only complement and might be take into account in developing the methods for diagnosing diseases. The basis of the rules of fuzzy ES work is the concept of linguistic variables. Each of them has a set of values - fuzzy variables that form its term set. The linguistic variable L is characterize by the following set of properties:

$$\mathrm{L} = (\mathrm{X}, \mathrm{T}(\mathrm{X}), \mathrm{U}, \mathrm{G}, \mathrm{M}), \tag{1}$$

where X is the name of the variable; $\mathrm{T}(\mathrm{X})$ - term-set of a variable X, that is, the set of names of the linguistic values of a variable X, each of these values being a fuzzy variable x' with values from a universal set U_3 with a base variable u; G is a syntactic rule that generates the names x' of the values of the variable X; M - a semantic rule that matches each fuzzy variable x' with its meaning $\mathrm{M}(\mathrm{x}')$, that is, the fuzzy subset $\mathrm{M}(\mathrm{x}')$ of the universal set U.

Fuzzy variable is characterized by $\langle \mathrm{x,U,X} \rangle$, where x - the name of the variable; U - universal set; X - fuzzy subset of the set U, which is a fuzzy constraint on the value of a variable $\mathrm{u} \in \mathrm{U}$, conditioned X.

It is proposed to analyze the data and knowledge that might be used for diagnosis. First, this study relies on a broad view of electronic medical data (EMD) about patients and illnesses.

It is also important to consider that for the problem under consideration in this study it is necessary to process a large amount of data. That is, Medical Big Date (MBD). However, for reliability of work of MES and reception of high-quality MKB it is necessary to analyze and provide high reliability of storage MKB.

In addition, it is necessary to develop the architecture and structure of the MES. This research direction relies on the development of UML models and diagrams that provide the development of an adaptive or self-organized MES. The software (software) of the MES must be developed on the basis of the use of modern MBD processing technologies such as OOP (Object Oriented Programming), Map Reduce, Hadoop, etc.

For the study of illness probability calculation, the Bayesian theorem is used - a theorem, which, based on circumstances, describes the probability of an event. In our example, we use its Bayesian interpretation, that is, the probability measures the measure of confidence. Bayesian theorem therefore links the degree of confidence in the statement before and after considering the testimony. There is described the formula of probability as follows:

$$P(C \mid E) = \frac{P(E \mid C) * P(E)}{P(C)}, \tag{2}$$

where $P(C \mid E)$ – probability that effect is caused;

$P(E\,|\,C)$ – probability that the effect will appear when the cause appears; $P(C)$ – probability of the cause; $P(E)$ – probability of the effect.

If the number of effects is greater than one, then the coincidence of the probability class are summed up:

$$P(C\,|\,E_1,E_2,...)=\frac{P(E_1\,|\,C)*P(E_1)+P(E_2\,|\,C)*P(E_2)+...}{P(C)}, \tag{3}$$

where $P(C\,|\,E_1,E_2,...)$ – probability that the cause causes this effect,

$P(E_{1,2,...}\,|\,C)$ – probability that the effect will appear when the cause appears,

$P(C)$ – probability of the cause,

$P(E_{1,2,...})$ – probability of the effect.

Thereof considering all causes as equiprobable, in order to find $P(E)$ and $P(C)$ the following formula is used:

$$P(E),P(C)=\frac{1}{n}, \tag{4}$$

where n is the number of elements of this type (for example, symptoms, or syndromes, etc.).

This way, you can always complete statistics when adding new data to the system by simply adding new information to the already calculated vertices of the nodes. Using Formula (3) as:

$$P(C\,|\,E_1,E_2,...,E_n)=P(C\,|\,E_1,E_2,...)+\frac{P(E_n\,|\,C)*P(E_n)}{P(C)}, \tag{5}$$

where $P(C\,|\,E_1,E_2,...)$ – preliminary probability of the node,

$\frac{P(E_n\,|\,C)*P(E_n)}{P(C)}$ – adding new information to the system,

$P(E_n\,|\,C)$ – probability that a new effect will appear when the cause appears,

$P(E_n)$ – probability of the effect.

Example of probability distribution

For example, we have input data (Table 2)

Table 2 – The values of a priori and conditional probabilities for hypotheses

$p(\ )/i$	1	2	3
$p(H_i)$	0,59	0,39	0,02
$p(E_1\,|\,H_i)$	0,49	0,89	0,39
$p(E_2\,|\,H_i)$	0,09	0,79	0,99

In this case, the initial hypothesis characterizes the event associated with determining the reliability of some disease:

- H_1 – "pneumonia";
- H_2 – "bronchitis";
- H_3 – "tuberculosis".

Events (conditionally independent evidence) that support the initial hypothesis are:

- E_1 – "high temperature";
- E_2 – "coughing".

In the process of gathering facts of probability, hypotheses will increase if the facts approve them or decrease if they disprove them. Suppose we have only one fact E_1 – "high temperature" (that is with probability equal to one has come a fact E_1 – "high temperature"). Having received E_1 – "high temperature" we compute the a posteriori probability for hypotheses according to the Bayesian formula for one fact:

$$p(H_i \mid E_1) = \frac{p(E_1 \mid H_i) * p(H_i)}{\sum_{k=1}^{3} p(E_1 \mid H_k) * p(H_k)}, \quad i = 1,2,3. \quad . \tag{6}$$

As follows:

- $p(H_1 \mid E_1) = \frac{0,49 * 0,59}{0,49 * 0,59 + 0,89 * 0,39 + 0,39 * 0,02} = 0,4489;$
- $p(H_2 \mid E_1) = \frac{0,89 * 0,39}{0,49 * 0,59 + 0,89 * 0,39 + 0,39 * 0,02} = 0,5390;$
- $p(H_3 \mid E_1) = \frac{0,39 * 0,02}{0,49 * 0,59 + 0,89 * 0,39 + 0,39 * 0,02} = 0,0121. \quad .$

We do a check, the sum of a posteriori probabilities as a result should give to one:

$$p(H_1 \mid E_1) + p(H_2 \mid E_1) + p(H_3 \mid E_1) = 1. \tag{7}$$

That is, $0,4489 + 0,5390 + 0,0121 = 1$.

After E_1 – "high temperature" the confidence in the hypothesis H_1 – "pneumonia" and H_3 – "tuberculosis" reduced, when it increased to H_2 – "bronchitis". In cases where there are facts confirming both event E_1 – "high temperature" and event E_2 – "coughing", then the a posteriori probability of the initial hypothesis can also be calculated by the Bayesian rule:

$$p(H_i \mid E_1E_2) = \frac{p(E_1E_2 \mid H_i) * p(H_i)}{\sum_{k=1}^{3} p(E_1E_2 \mid H_k) * p(H_k)}, \quad i = 1,2,3. \tag{8}$$

Because of conditional independence of high temperature and coughing, we can rewrite Bayesian formula as:

$$p(H_i \mid E_1E_2) = \frac{p(E_1 \mid H_i) * p(E_2 \mid H_i) * p(H_i)}{\sum_{k=1}^{3} p(E_1 \mid H_k) * p(E_2 \mid H_i) * p(H_k)}, \quad i = 1,2,3. \tag{9}$$

As follows:

- $p(H_1 \mid E_1E_2) = \frac{0{,}49 * 0{,}09 * 0{,}59}{0{,}49 * 0{,}09 * 0{,}59 + 0{,}89 * 0{,}79 * 0{,}39 + 0{,}39 * 0{,}99 * 0{,}02} = 0{,}0845;$
- $p(H_2 \mid E_1E_2) = \frac{0{,}89 * 0{,}79 * 0{,}39}{0{,}49 * 0{,}09 * 0{,}59 + 0{,}89 * 0{,}79 * 0{,}39 + 0{,}39 * 0{,}99 * 0{,}02} = 0{,}8904;$
- $p(H_3 \mid E_1E_2) = \frac{0{,}39 * 0{,}99 * 0{,}02}{0{,}49 * 0{,}09 * 0{,}59 + 0{,}89 * 0{,}79 * 0{,}39 + 0{,}39 * 0{,}99 * 0{,}02} = 0{,}0251.$

Equivalence probability check:

$$p(H_1 \mid E_1E_2) + p(H_2 \mid E_1E_2) + p(H_3 \mid E_1E_2) = 1. \tag{10}$$

That is, $0{,}0845 + 0{,}8904 + 0{,}0251 = 1$.

Initial ranking was H_1 – "pneumonia", H_2 – "bronchitis" and H_3 – "tuberculosis", and all three remained after receiving the facts E_1 – "high temperature" and E_2 – "coughing". Herewith bronchitis more likely than pneumonia and tuberculosis. This indicates that having coughing and high temperature the probability of the disease bronchitis much bigger than the probability of the disease pneumonia or tuberculosis.

However, realistically, the spread of probabilities occurs in stages with the summation of individual facts and their impact on the probabilistic probability of the receiving the individual values E_i. It proceeds by using the a priori and a posteriori probability, thus:

1. Define $p(H_i)$ – a priori probability of events H_i.
2. For the received facts E_i set down $p(E_i \mid H_i)$.
3. Considering the Bayesian theorem calculate $p(H_i \mid E_i)$ depending on the outcome E_i, that is, we calculate the a posteriori probability of the event H_i.

4. Now you can mark the current a posteriori event probability H_i, as a new a priori probability Hi. Therefore, $p(H_i)$ equals $p(H_i \mid E_i)$ depending on the value E_i

5. Then choose a new fact for consideration and proceed to step two.

Consider an example, the fact E_2 – "coughing" entered the system. Then:

– $p(H_1 \mid E_2) = \dfrac{0{,}09 * 0{,}59}{0{,}09 * 0{,}59 + 0{,}79 * 0{,}39 + 0{,}99 * 0{,}02} = 0{,}1394;$

– $p(H_2 \mid E_2) = \dfrac{0{,}79 * 0{,}39}{0{,}09 * 0{,}59 + 0{,}79 * 0{,}39 + 0{,}99 * 0{,}02} = 0{,}8087;;$

– $p(H_3 \mid E_2) = \dfrac{0{,}99 * 0{,}02}{0{,}09 * 0{,}59 + 0{,}79 * 0{,}39 + 0{,}99 * 0{,}02} = 0{,}0519.$

Check:

$$p(H_1 \mid E_2) + p(H_2 \mid E_2) + p(H_3 \mid E_2) = 1. \quad (11)$$

That is, $0{,}1394 + 0{,}8087 + 0{,}0519 = 1.$

We take the resulting probability as a new a posteriori probability of hypothesis H_1, H_2 and H_3, so:

– $p\left(\tilde{H}_1\right) = 0{,}1394;$

– $p\left(\tilde{H}_2\right) = 0{,}8087;$

– $p\left(\tilde{H}_3\right) = 0{,}0519.$

Now, if any additional fact like E_1 – "high temperature" arrives, then the new a posteriori probability of the hypothesis calculates only on the evidence that arrives again:

– $p(H_1 \mid E_1 E_2) = p\left(\tilde{H}_1 \mid E_1\right) = \dfrac{0{,}49 * 0{,}1394}{049 * 0{,}1394 + 0{,}89 * 0{,}8087 + 0{,}39 * 0{,}0519} = 0{,}0845;$

– $p(H_2 \mid E_1 E_2) = p\left(\tilde{H}_2 \mid E_1\right) = \dfrac{0{,}89 * 0{,}8087}{049 * 0{,}1394 + 0{,}89 * 0{,}8087 + 0{,}39 * 0{,}0519} = 0{,}8905;$

– $p(H_3 \mid E_1 E_2) = p\left(\tilde{H}_3 \mid E_1\right) = \dfrac{0{,}39 * 0{,}0519}{049 * 0{,}1394 + 0{,}89 * 0{,}8087 + 0{,}39 * 0{,}0519} = 0{,}0250.\ .$

Check:

$$p\left(\tilde{H}1 \mid E_1\right) + p\left(\tilde{H}2 \mid E_1\right) + p\left(\tilde{H}3 \mid E_1\right) = 1. \quad (12)$$

That is, $0{,}0845 + 0{,}8905 + 0{,}0250 = 1.$

From the example given, you can see that the iterative procedure for the sequential probability distribution during the receiving of new facts allows you to get the results similar to the results of the Bayesian rule for two simultaneously received facts. The value of the hypothesis H_2 – "bronchitis" is most likely then H_1 – "pneumonia" and H_3 – "tuberculosis".

Conclusion

There are ES, built on objective and subjective views on the concept of the probability of an event. Knowledge bases of the ES accumulate human knowledge. Therefore, interpretations based on subjective trust are best suited to represent experts' knowledge in the light of probabilities. As a result, most of the modern EUs who use the theory of probabilities are "Bayesian".

The use of the Bayesian strategy in the ES implements with using the Bayesian formula of "inverse probabilities", that is, the estimation of conditional probabilities of hypotheses. In the presence of several signs (symptoms), such calculations are simply realized in the assumption of statistical independence of the characteristics, which is far from always corresponds to reality. However, practice shows that such an approach, despite its obvious mathematical incorrectness, is quite applicable, since it usually leads to correct conclusions.

Expert systems that use the theory of subjective probabilities are widely used in medicine as well as in other areas where it is necessary to determine the probability of occurrence of a certain event clearly and meaningfully. The theory of subjective probabilities subordinate directly to the Bayesian theory. It is used to evaluate a specific task, analyzing it, giving a solid answer, and making predictions for the future.

During the calculations of the distribution of probabilities in expert systems with given hypotheses, H_1 - "pneumonia", H_2 - "bronchitis", H_3 - "tuberculosis", characterizing the event associated with the definition of some disease, the result was obtained, indicating the probability the appearance of bronchitis, more than two other specified diseases.

Bibliography

[*Oleksandr Kuzomin. 2017*] DEVELOPMENT OF INTELECTUAL MODELS AND METHODS OF EXPERT SYSTEMS OF CLINICAL MEDICINE.// Oleksya Vasilenko, Tatyana Tolmachova International Journal “Informattion Technologes&Knolegedge”. Volume 11, Namber 2, 2017. PP. 186-199 ISSN 1313-0455 (printed), ISSN 1313-048X (online), [*Oleksandr Kuzomin. 2014*] Data loss minimization in situation's centrums databases // Oleksandr Ya. Kuzomin, Oleksii Vasylenko. Chairman IDRC Davos 2014 - Global Risk Forum GRF Davos - Davos – Switzerland. PP 153-154.

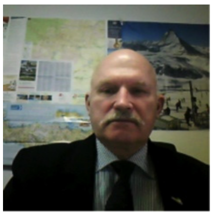
[*Oleksandr Kuzomin. 2017*] METHODS AND MODELS FOR BUILDING A DISTRIBUTED MOBILE EMERGENCY MONITORING SYSTEM.Oleksandr Kuzomin, Oleksii. Vasylenko, 17th International Multidisciplinary Scientific Geoconference SGEM 2017. Conference Proceedings, Informatics Geoinformatics. Volume 17. ISSUE 21 PP 433 – 440. ISBN 978-619-7408-01-0, ISSN 1314-2704. DOI: 10.5593/sgem2017/2.1,

Authors' Information

***Oleksii Vasylenko** – Aspirant of Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;*

***e-mail**: ichbierste@gmail.com tel.: +380 63 841 66 23*

***Major Fields of Scientific Research**: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics.*



***Prof. Dr. Oleksandr Kuzomin** – Informatics chair Innovation-marketing department of Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;*

***e-mail**: kuzy@daad-alumni.de tel.: +38(057)7021515*

***Major Fields of Scientific Research**: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.*

***Oleksandr Shapoval** – Master student in Information and Communication technologies of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine*

***Major Fields of Scientific Research**: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics.*