УДК 61:001.8
DOI: http://dx.doi.Org/10.11603/mie.1996-1960.2016.4.7055

ON PREDICTION AND DECISION MAKING UNDER UNCERTAINTIES FOR MEDICAL SYSTEMS RESEARCH PROBLEMS

V. P. Martsenyuk , I. Ye. Andrushchak[1]

University of Bielsko-Biala, the Republic of Poland
[1]Lutsk National Technical University

The work presents our results in field of application of system analysis methods to problem of medical research. We emphasize effects of uncertainty that should be taken into account in such complex processes. Medical system research requires information support system implementing data mining algorithms resulting in decision trees or IF-THEN rules. Besides that such system should be object-oriented and web-integrated.

The aim of this study was to develop information support system based on data mining algorithms applied to system analysis method for medical system research. System analysis methods were used for qualitative analysis of diseases mathematical models. Algorithms such as decision tree induction and sequential covering algorithm were applied for data mining from learning data set.

We observed the complex qualitative behavior of population and diseases models depending on parameters and controllers even without considering probabilistic nature of the most of quantities and parameters of information models.

Key words: system analysis, decision support systems, information system, simulation, optimization, dynamic system, qualitative analysis, decision tree, classification rule, health research systems.

ПРОГНОЗУВАННЯ ТА ПРИЙНЯТТЯ РІШЕНЬ В УМОВАХ НЕВИЗНАЧЕНОСТІ В ПРОБЛЕМАХ СИСТЕМНИХ МЕДИЧНИХ ДОСЛІДЖЕНЬ

В. П. Марценюк, І. Є. Андрущак[1]

Університет Бєльсько-Бяли, Республіка Польща
[1]Луцький національний технічний університет

Робота представляє наші результати в галузі застосування методів системного аналізу до проблеми медичних наукових досліджень. Акцент робиться на ефектах невизначеності, які слід брати до уваги в таких складних процесах. Медичні системні дослідження потребують системи інформаційної підтримки, що реалізує алгоритми data mining з побудовою дерев рішень та правил типу IF-THEN. До того ж така система повинна бути об'єктно-орієнтованою і веб-інтегрованою.

Метою цього дослідження є розробка системи інформаційної підтримки, яка ґрунтується на алгоритмах data mining, що застосовуються в методах системного аналізу системних медичних досліджень. Методи системного аналізу застосовуються для якісного аналізу математичних моделей захворювань. Алгоритми, такі як індукція дерева рішень та алгоритм послідовного покриття, застосовуються для data mining, виходячи з навчальних наборів даних.

Спостерігається складна якісна поведінка моделей популяцій та захворювань, що залежить від параметрів та керування, навіть без врахування ймовірнісної природи більшості величин і параметрів інформаційних моделей.

Ключові слова: системний аналіз, системи прийняття рішень, інформаційна система, моделювання, оптимізація, динамічна система, якісний аналіз, дерево рішень, класифікаційне правило, медичні наукові системи.

ПРОГНОЗИРОВАНИЕ И ПРИНЯТИЕ РЕШЕНИЙ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ В ПРОБЛЕМАХ СИСТЕМНЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

В. П. Марценюк, И. Е. Андрущак[1]

Университет Бельско-Бялы, Республика Польша
[1]Луцкий национальный технический университет

Работа представляет наши результаты в области применения методов системного анализа к проблеме медицинских научных исследований. Акцент делается на эффектах неопределенности, которые следует принимать во внимание в таких сложных процессах. Медицинские системные исследования требуют системы информационной поддержки, которая реализует алгоритмы data mining с построением деревьев решений и правил типа IF-THEN. К тому же такая система должна быть объектно-ориентированной и веб-интегрированной.

Целью настоящего исследования является разработка системы информационной поддержки, основанной на алгоритмах data mining, применяемых в методах системного анализа системных медицинских исследований. Методы системного анализа применяются для анализа математических моделей заболеваний. Алгоритмы, такие как индукция дерева решений и алгоритм последовательного покрытия, применяются для data mining, исходя из учебных наборов данных.

Наблюдается сложное качественное поведение моделей популяций и заболеваний, которое зависит от параметров и управления, даже без учета вероятностной природы большинства величин и параметров информационных моделей.

Ключевые слова: системный анализ, системы принятия решений, информационная система, моделирование, оптимизация, динамическая система, качественный анализ, дерево решений, классификационное правило, медицинские научные системы.

Introduction. Here we would like to present our results in field of application of system analysis methods to problems of medical science. We emphasize effects of uncertainty that should be taken into account in such complex medical systems. It will be shown that even considering deterministic models of such nonlinear systems we see different qualitative behavior closely dealt with parameters values.

Let's start from origin of such a problem. Nowadays there are obtained a lot of models describing physiological indices of human body at different diseases and treatment schemes. Primarily they are based on regression analysis. More complex ones use neural networks and evolutionary programming. The most significant attempts to construct mathematical models at different levels of hierarchy of human organism were made by John Murray [8], Keener and Sneyd [1], G. I. Marchuk [2], Mackey and Glass (they investigated nonlinear phenomena applying dynamic systems and introduced notion of dynamic diseases). Without considering uncertainty all these models can be applied for patients from determined groups (primarily for given age and a lot of another restrictions).

As for projects stimulating given research we would like to note the following. During the last decade we are fulfilling investigations initiated by Healthcare Ministry of the Ukraine in order to develop and use general system analysis algorithm to study different diseases [3-7, 9]. Namely, in fields of oncology (melanoma, leukemia), infectious diseases (flue), therapy (bone tissue diseases). Naturally there arises a problem to develop a general model for disease. It is incorrect to state that we managed offering unique universal algorithm to construct disease general model at whole. More correct is to say this approach can be used for diseases of different nature. We believe this approach can be extended to processes in sociology and demography as well as for economy and finance branches tasks. A lot of them have the same nature as human diseases. Let's pay attention on special medical terminology necessary (as small as possible). First of all, the most recognized definition of disease states that disease is a set of pathologic processes weakening vitality and activity of a human organism. Here pathologic process is a set of pathologic (that is not normal) and protectoral reactions within human organism. That is, the most significant is modeling pathologic process.

Based on this reasoning we offered general model for pathologic process including three counterparts:

- the reason or cause of disease (it may be some external factor (like bacteria, chemicals) or own modified cells (tumor cells);

- immune system supports organism with help of specific antibodies (sort of predators) and plasmatic cells (their ancestors);
- normal cells, tissues and organs (it is necessary to consider them to satisfy to some constraints of toxicity).

For these researches we used our own software - Software Environment for Medical System Researches (SEMSR). There is developed conceptual model of software environment of system medical investigations support. Implementing it there is offered model of data structure in branch of system medical investigations and invented in terms of XML-technology. There is developed interface which is Web-integrated, user-oriented and adjustable. There are implemented mathematical methods of system analysis of pathologic processes in form of Java-classes hierarchy. There are developed software tools to execute system medical investigations, to prepare results obtained for presentation in Internet and visualization.

Uncertainties in medical system research. Uncertainties in such models may be parametric. Some of the parameters may be unknown functions. As for uncertainty in control it is necessary to take into account all possible scenarios. Note, the purpose of this article is not to present methods to identify these uncertainties. For these purpose we need to present powerful and deep mathematical apparatus of adjoint systems, sensitivity functions and minimax aposteriorial estimation. Here we would like to answer two questions.

- Why is it so important to take into account uncertainties?
- The basic uncertainties in models of diseases.

When answering the first question we should say that as it was shown even mathematical solutions of equations have different qualitative behavior. In practice we can observe different forms of disease (subclinical, acute, chronic, and lethal). Search of treatment scheme is dependent on such forms.

In our research we investigated uncertainties in the following issues: maturation time for plasmatic cells **T**, influence of antigen on target-organ damage rate o, relation between target-organ damage rate and immune response ^ (m), therapy scheme (polychemiotherapy, radiotherapy), surgery interventions. Note, the three last ones are non-parametric. They depend on unknown function like controller.

Approach of compartmental systems. Problems of population dynamics, pharmacokinetics, mathematical epidemiology, and others are described by compartmental systems with time delay. Even in the linear case, the solution of such equations leads to approximate computation procedures, which makes it impossible to find solutions of the following problems in explicit form:

- determining the time instant at which the number of infected persons does not exceed some level i* (mathematical epidemiology);
- estimating the time when no more than d* medical product units (pharmacokinetics) remain in the organism of a patient, etc.

Explicit solutions of such problems can be obtained on the basis of exponential type estimates. A number of works are devoted to the construction of exponential estimates for systems with delay. In particular, in [2], an estimate for a linear system is obtained on the basis of the Cauchy formula. An approach based on Lyapunov functions with conditions of the Razumikhin type was developed in [1]. In [8], an estimate is found from the solution of a difference inequality for a Lyapunov - Krasovskii functional. In [5], a differential difference inequality is constructed for a Lyapunov - Krasovskii functional. For compartmental systems, a promising approach is proposed in [3] in which the method of construction

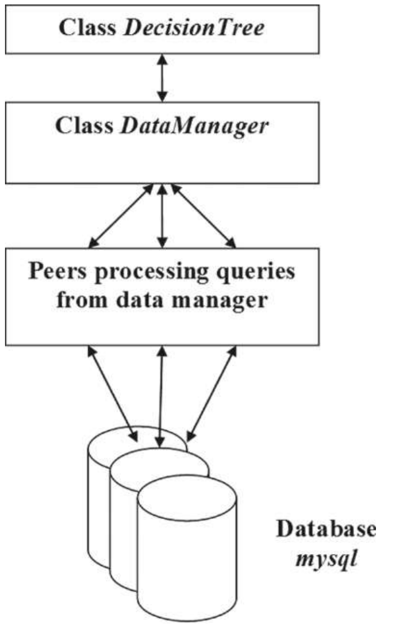

Fig. 1. Conceptual model of informational system of decision tree induction

of a class of exponential estimates is based on the Hale - Lunel inequality.

Software development based on data mining technology. The objective is to develop and implement algorithms of diagnostic classification applying decision tree induction and sequential covering methods and to study problem of their computational complexity.

The problem solved belongs to wide class of differential diagnostics problems. In medicine the notion of "differential diagnostics" means systemic approach based on evidence for determining causes of symptoms observed in case if there are few alternative explanations and also to reduce list of possible diagnoses.

One of approaches expressing natural process of thinking for differential diagnostics is data mining method. We are interested in the problem of computational complexity of the algorithms for real clinical data such as, for example, for biochemical data in case of polytraumas.

Software implementation of decision tree induction. The methods are implemented within Netbeans developer system in Java language. The database of learning tuples is deployed on MySQL server. In fig. 1 there is presented conceptual model of informational system. Class *DecisionTree* implements decision tree induction method. Class *DataManager* is processing calls from *DecisionTree* running queries to *mysql* database retrieving learning data.

Database *mysql* consists of two tables - table *attribute* for storage of information on attributes and table *categorized_data* - for learning tuples. The structure of tables in SQL syntax is shown below:

```
CREATE TABLE mysql.attribute (
id integer not null unique,
            attribute_name     varchar(25),
            attribute_field_name     varchar(25),
      primary key (id)
 ) ENGINE=InnoDB;
CREATE TABLE mysql.categorisedjdata (
id integer not null unique,
            A1   varchar(12),
            A2   varchar(8),
            A3   varchar(7),

            A21   varchar(7),
            class   varchar(28),
      primary key (id)
 ) ENGINE=InnoDB;
```

Classes of this project are included in package *decisionjtree.model.* Here there are beans-classes *Attribute, Attribute_for_list* and *CategorisedData* for processing data of corresponding tables. SQL-queries for retrieving corresponding data including calculations of information indices are implemented in class *AttributeListPeer.*

Problem of computational complexity of decision tree induction algorithm. As it was shown in the work [Han, 2001] time of decision tree induction algorithm running is estimated with value

$$O(p\times\#(D)\times\log(\#(D))) \qquad (1)$$

Our goal was to check this result experimentally. Experiments were executed varying amount of attributes *p*. Decision trees were constructed for each value of *p*. In fig. 2 and 3 there are shown estimates of decision tree induction times due to (1).

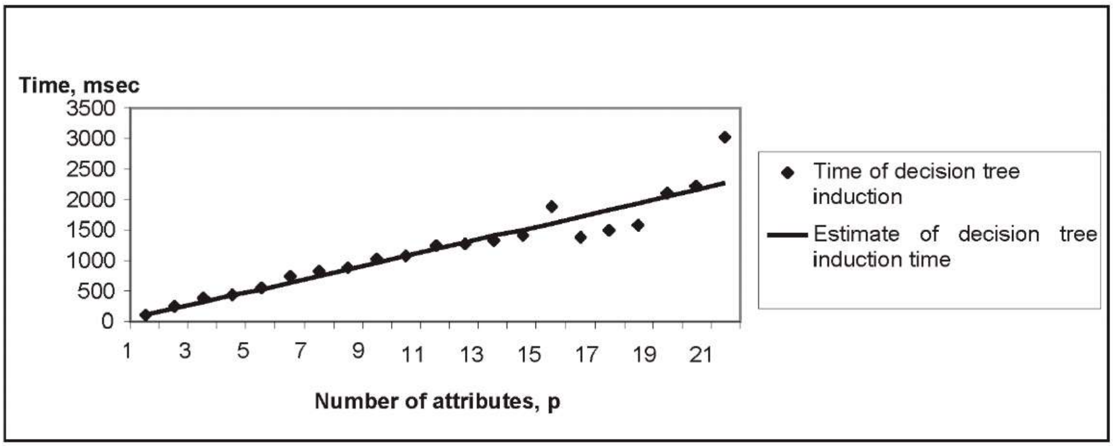

Fig. 2. Estimate of algorithm complexity based on information gain

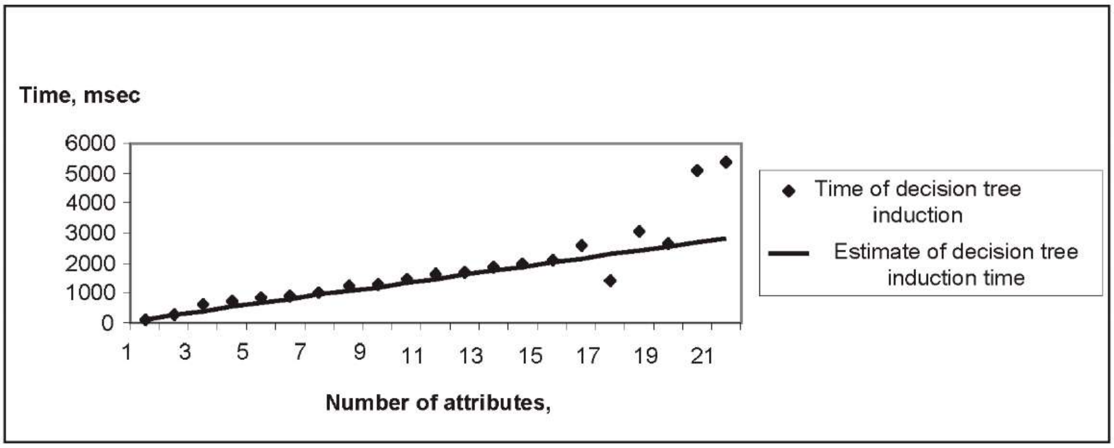

Fig. 3. Estimate of complexity based on information gain ratio

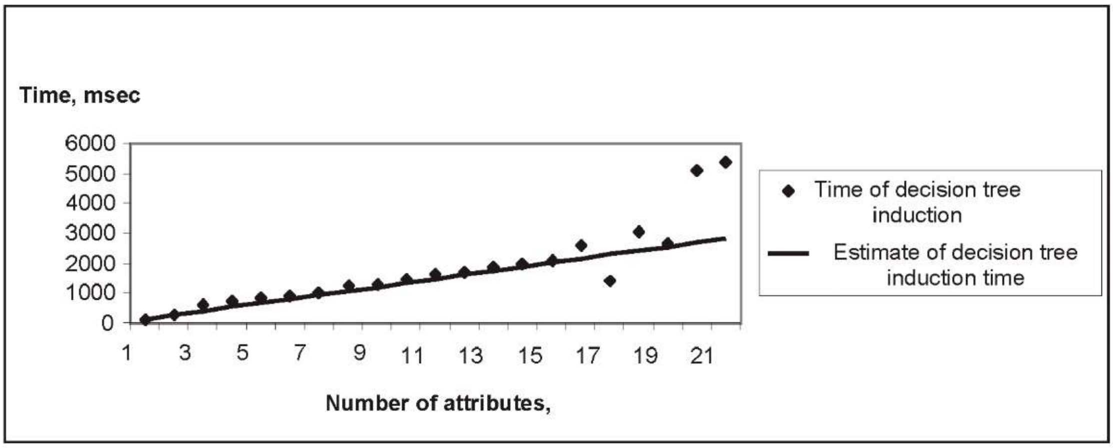

Fig. 4. Estimate of complexity of sequential covering algorithm

Computational complexity of sequential covering algorithm. Due to analysis of sequential covering algorithm we conclude that computational complexity is determined by product of amount of possible values of class attribute *K* (quantity of external cycle iterations) and computational complexity of procedure *Mine_one_rule (D, Att_vals, c)* executed inside each

Procedure *Mine_one_rule (D, Att_vals, c)* includes execution of*p* iterations. For each iteration for a certain attribute A_i we calculate the measure *FOIL_Gain* for each of K_i values of attribute. That is internal body of cycle in procedure *Mine_one_rule (D, Att_vals, c)* is executed times

$$\sum_{i=1}^{p} K_i$$

The measure *FOIL_Gain* is executed as a result of 4 SQL-queries with complexity O(log(N)) (according with MySQL 5.0 documentation). That is procedure *Mine_one_rule (D, Att_vals, c)* has computational complexity

$$O\left(\sum_{i=1}^{p} K_i \times \log(N)\right)$$

Summarizing we have sequential covering algorithm complexity of the order

$$O\left(K \times \sum_{i=1}^{p} K_i \times \log(N)\right) \qquad (2)$$

In fig. 4 there is shown estimates of sequential covering algorithm times due to (2).

Conclusions. So, even without considering probabilistic nature of the most of quantities and parameters we saw the complex qualitative behavior of diseases models depending on parameters and controllers. At different values of these quantities we observed subclinical, acute, chronic or lethal forms of pathologic processes.

Taking into account complexity of mathematical equations (nonlinear systems with delays) requires appearance of new powerful methods of exact parameter identification and qualitative analysis.

From viewpoint of theoretical medicine uncertainties arising in models of diseases require to develop treatment schemes that are effective, take into account toxicity constraints, enable life quality, cost benefit.

In future works our idea is to compare behavior of pathologic processes using both deterministic and stochastic models and to extend such models to demographic processes.

In the work there is considered the problem of development and implementation of decision tree induction and sequential covering methods based on information indices for construction of diagnostic classification algorithm.

When investigating in this example the problem of computational complexity of decision tree induction algorithm it was observed that:

- decision tree induction time based on information indices is well approximated with estimate (1) at small number of attributes (in this case to 15-16);
- when increasing number of attributes (in this example over 15-16) the time of decision tree induction begins deviate essentially from estimate (1) independent on search of information measure;
- at small number of attributes decision trees induced constructed based on either information gain or information gain are identical; i. e., information measure determining splitting attribute doesn't affect on decision tree induced;
- computational complexity of sequential covering algorithm is well approximated by (2). Such estimate was checked changing an amount of attributes as well as number of learning tuples.

The perspective of this investigation is comparative performance analysis depending on volume of set of learning tuples.

Література.

1. Keener J. Mathematical Physiology / J. Keener, J. Sneyd. - New York : Springer Verlag, 1998. - 768 p.
2. Mathematical modelling in immunology and medicine / Ed. by G. I. Marchuk, L. N. Belykh // IFIP TC-7 Working Conf. (5-11 July, 1982, Moscow). - Amsterdam, New York, Oxford : North-Holland, 1983.
3. Martsenyuk V. P. On Hopf bifurcation and periodic solutions in G. I. Marchuk model of immune protection / V. P. Martsenyuk // Journal of Automation and Information Sciences. - 2003. - Vol. 35, No. 8.
4. Martsenyuk V. P. On stability of immune protection model with regard for damage of target organ: the degenerate Lyapunov functionals method / V. P. Martsenyuk // Cybernetics and Systems Analysis. - 2004. - Vol. 40, No. 1. - P. 126-136.
5. Martsenyuk V. P. On the problem of chemotherapy scheme search based on control theory / V. P. Martsenyuk // Journal of Automation and Information Sciences. - 2003. -Vol. 35, No. 4.
6. Martsenyuk V. P. Qualitative analysis of human cells dynamics: stability, periodicity, bifurcations, control problems /V. P. Martsenyuk //Adv. Math. Res. - 2003. - Vol. 5, No. 1. - P. 137-200.
7. Martsenyuk V. P. Taking into account delay in the problem of immune protection of organism / V. P. Martsenyuk // Nonlinear Analysis: Real World Applications. - 2001. - Vol. 2, No. 4. - P. 483^196.
8. Murray J. M. Mathematical Biology / J. M. Murray. - New York : Springer-Verlag, 1989. - 768 p.
9. NakonechnyiA. G. Controllability problems for differential Gompertzian dynamic equations / A. G. Nakonechnyi, V. P. Martsenyuk // Cybernetics and Systems Analysis. - 2004. - Vol. 40, No. 2. - P. 252-259.

References.

1. Keener, J., Sneyd, J. (1998). Mathematical Physiology. New York: Springer Verlag.
2. Mathematical modelling in immunology and medicine : Proc. of the IFIP TC-7 Working Conf., Moscow, USSR, 5-11 July 1982, Ed. by G. I. Marchuk, L. N. Belykh. Amsterdam, New York, Oxford: North-Holland, 1983.
3. Martsenyuk, V. P. (2003). On Hopf bifurcation and periodic solutions in G. I. Marchuk model of immune protection. Journal of Automation and Information Sciences, 35(8). doi: 10.1615/JAutomatInfScien.v35.i8.70.
4. Martsenyuk, V. P. (2004). On stability of immune protection model with regard for damage of target organ: the degenerate Lyapunov functionals method. Cybernetics and Systems Analysis, 40(1), 126-136. doi: 10.1023/B:CASA.0000028109.69242.38.
5. Martsenyuk, V. P. (2003). On the problem of chemotherapy scheme search based on control theory. Journal of Automation and Information Sciences, 35(4). doi: 10.1615/JAutomatInfScien.v35.i4.60.
6. Martsenyuk, V. P. (2003). Qualitative analysis of human cells dynamics: stability, periodicity, bifurcations, control problems. Adv. Math. Res., 5(1), 137-200.
7. Martsenyuk, V. P. (2001). Taking into account delay in the problem of immune protection of organism. Nonlinear Analysis: Real World Applications, 2(4), 483-496. doi: 10.1016/S 1468-1218(01)00005-0.
8. Murray, J. M. (1989). Mathematical Biology. New York: Springer-Verlag.
9. Nakonechnyi, A. G., Martsenyuk, V. P. (2004). Controllability problems for differential Gompertzian dynamic equations. Cybernetics and Systems Analysis, 40(2), 252-259. doi:10.1023/B:CASA.0000034451.73657.88.