International Journal of Medicine and Medical Research
2015, Volume 1, Number 1, p. 63-67

INFORMATION SUPPORT SYSTEM OF MEDICAL SYSTEM RESEARCH

V. P. Martsenyuk[1], I. Ye. Andrushchak[2]
[1]TERNOPIL STATE MEDICAL UNIVERSITY, TERNOPIL, UKRAINE
[2]LUTSK NATIONAL TECHNICAL UNIVERSITY, LUTSK, UKRAINE

Background. *Medical system research requires information support system of implementing data mining algorithms resulting in decision trees or IF-THEN rules. Besides that, this system should be object-oriented and web-integrated.*

Objective. *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.*

Methods. *System analysis methods are used for qualitative analysis of mathematical models diseases. Algorithms such as decision tree induction and sequential covering algorithm are applied for data mining from learning data set.*

Results. *Taking into consideration the complexity of mathematical equations (nonlinear systems with delays), scientific community requires the appearance of new powerfull methods of exact parameter identification and qualitative analysis. From the point of view of theoretical medicine, uncertainties arising in models of diseases require to develop treatment schemes that are effective, take into account toxicity constraints, enable better life quality, have cost benefit. Multivariate method of qualitative analysis of mathematical models can be used for pathologic process forms of classification.*

Conclusions. *The complex qualitative behavior of diseases models depending on parameters and controllers was observed in our investigation even without considering probabilistic nature of the majority of quantities and parameters of information models.*

KEY WORDS: **data mining, system analysis, medical research, decision making**

Introduction

Here, we would like to present our results in field of application of system analysis methods to problem of clinical medicine. We emphasize that effects of uncertainty should be taken into account in such complex systems. It will be shown that even considering deterministic models of such nonlinear systems, we observe different qualitative behavior closely dealt with parameters values. Let's start from the origin of this problem. Nowadays, a lot of models describing physiological indices of human body at different diseases and treatment schemes are obtained. 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 [3], Keener and Sneyd [2], G.I.Marchuk [1], 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 other restrictions).

Address for correspondence: Vasil Martsenyuk, I. Ya. Horbachevsky Ternopil State Medical University, m. Voli, 1, Ternopil, 46001, Ukraine
E-mail: marceniuk@yahoo.com

Methods

As for projects stimulating given research, we would like to note the following. During the last years Medical Informatics Department performs investigations initiated by Healthcare Ministry of Ukraine in order to develop and use general system analysis algorithm to study different diseases [4–9]. Namely, in fields of oncology (melanoma, leukemia), infectious diseases (flu), therapy (bone tissue diseases). Naturally, there arises a problem to develop a general model for disease. It is incorrect to state that we managed to offer unique universal algorithm to construct disease general model. More correctly is to say that 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. A lot of them have the same nature as human diseases. Let's take into consideration special medical terminology (as little 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 abnormal) and protectoral reactions within human organism. The most significant is modeling pathologic process.

Results

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

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

(ii) *immune system* supports organism with help of specific antibodies (sort of predators) and plasmatic cells (their ancestors);

(iii) *normal cells*, tissues and organs (it is necessary to consider them to satisfy some constraints of toxicity).

We used our own software for these researches: Software Environment for Medical System Researches (SEMSR). Conceptual model of software environment of system medical investigations support is developed. Model implementation of data structure for medical investigations in terms of XML-technology is offered. Interface which is Web-integrated, user-oriented and adjustable is developed. Mathematical methods of system analysis of pathologic processes in form of Java-classes hierarchy are implemented. Software tools to execute system medical investigations, to prepare results obtained for presentation in Internet and visualization are developed.

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:

(i) why is it so important to take into account uncertainties?

(ii) the basic uncertainties in models of diseases.

To answer question № 1, we should say that the mathematical solutions of equations have different qualitative behavior. In practice we can observe different forms of disease (subclinical, acute, chronic, 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 τ, influence of antigen on target-organ damage rate σ, relation between target-organ damage rate and immune response $\xi(m)$, 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 [1], 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 [2]. In [3], an estimate is found from the solution of a difference inequality for a Lyapunov–Krasovskii functional. In [4], a differential difference inequality is constructed for a Lyapunov–Krasovskii functional. For compartmental systems, a promising approach is proposed in [5] and the method of construction 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 an algorithms of diagnostic classification applying decision tree induction and sequential covering methods and to study problem of their computational complexity.

The solved problem 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 a 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. At fig.1 the conceptual model of informational system is presented. 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.categorised_data (
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 *decision_tree.model*. 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 [11], time of decision tree induction algorithm running is estimated with value

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

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. At fig. 2 and 3 there are shown estimates of decision tree induction times according to [4].

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 cycle.

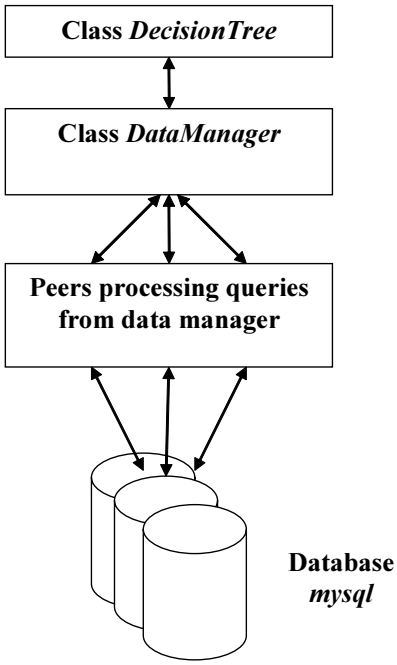

Fig. 1. Conceptual model of informational system of decision tree induction

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 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 $\sum_{i=1}^{p} K_i$ times. The measure 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

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

Conclusions

So, even without considering probabilistic nature of the majority 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 consideration complexity of mathematical equations (nonlinear systems with delays),

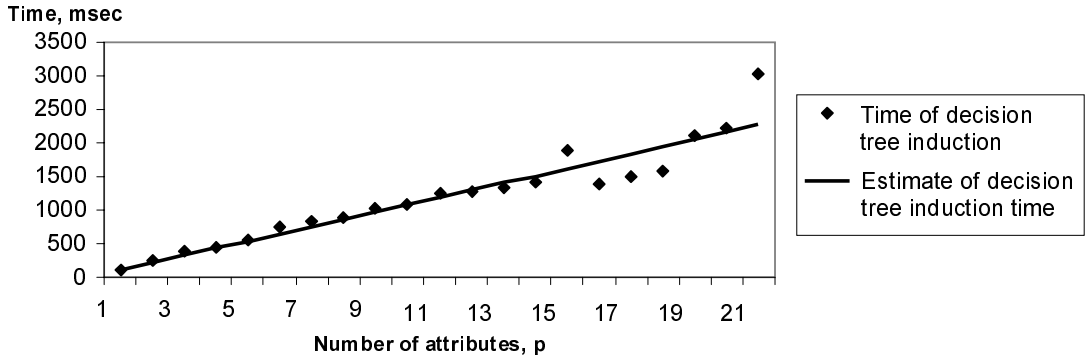

Fig. 2. Estimate of algorithm complexity based on information gain.

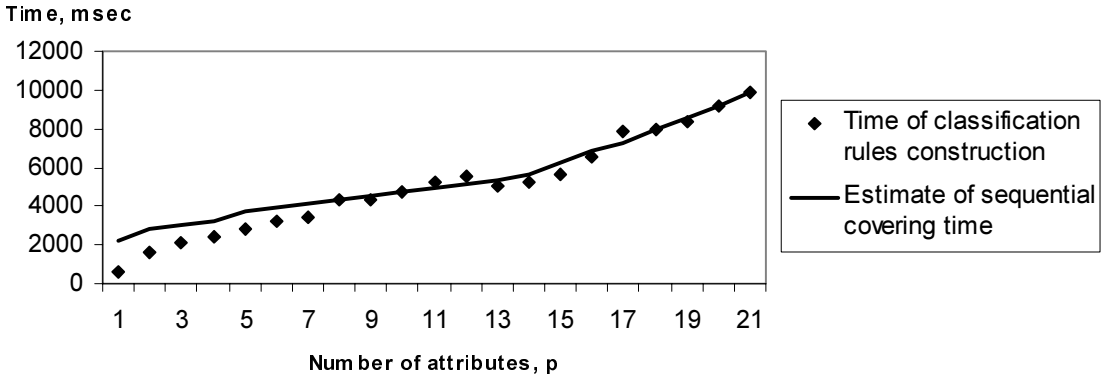

Fig. 3. Estimate of complexity of sequential covering algorithm.

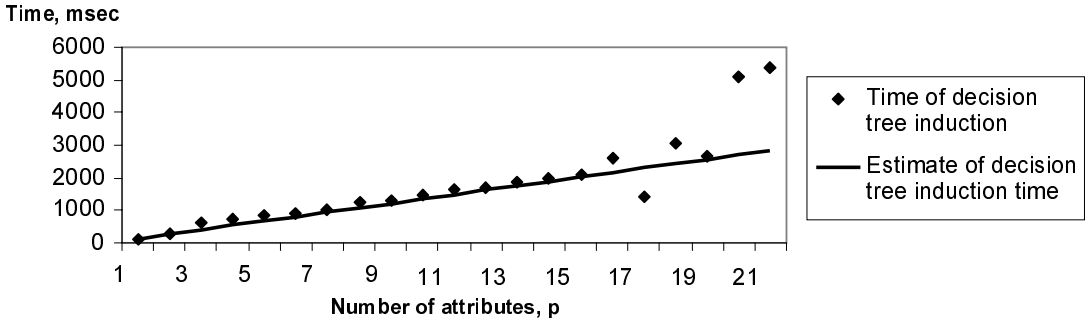

Fig. 4. Estimate of complexity based on information gain ratio.

it requires appearance of new powerfull methods of exact parameter identification and qualitative analysis.

From point of view of theoretical medicine, uncertainties arising in models of diseases require development of treatment schemes that are effective, take into account toxicity constraints, enable better life quality, have cost benefit.

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

In the work here we 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.

While investigating this example, the problem of computational complexity of decision tree induction algorithm 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 to 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 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.

References

1. 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: 246.

2. Keener J, Sneyd J. Mathematical physiology. New York: Springer Verlag; 1998: 149.

3. Murray JM. Mathematical biology. New York: Springer-Verlag; 1989: 214.

4. Martsenyuk VP. On the problem of chemotherapy scheme search based on control theory. J Automation Information Sci 2003; 35 (4): 64–69.

5. Martsenyuk VP. On Hopf bifurcation and periodic solutions in G.I. Marchuk model of immune protection. J Automation Information Sci 2003; 35 (8): 154–157

6. Marzeniuk VP. Taking into account delay in the problem of immune protection of organism. Nonlinear Analysis: Real World Applications 2001; 2 (4): 483–496.

7. Nakonechnyi AG, Martsenyuk VP. Controllability problems for differential gompertzian dynamic equations. Cybernetics Systems Analysis 2004; 40 (2): 252–259.

8. Martsenyuk VP. On stability of immune protection model with regard for damage of target organ: the degenerate Liapunov functionals method. Cybernetics Systems Analysis 2004; 40 (1): 126–136.

9. Marzeniuk VP. Qualitative analysis of human cells dynamics: stability, periodicity, bifurcations, control problems. Adv Math Res 2003; 1 (5): 137–200.

10. Khusainov DYa, Martsenyuk VP. Double-ended estimates for solutions of linear systems with delay. Dop. NAN Ukr 1996; 8: 8–13.

11. Han J, Kamber M. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 1st edition; 2001: 312.

12. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning, Springer, New York, 1st edition; 2001: 125.

13. Ordonez C., Comparing association rules and decision trees for disease prediction. In Proc. ACM HIKM Workshop 2006: 17–24.

14. Ordonez C. Integrating K-means clustering with a relational DBMS using SQL, IEEE Transactions on Knowledge and Data Engineering (TKDE) 2006; 18(2): 188–201.

15. Quinlan JR. Induction of decision trees. Machine Learning 1986; 1: 81–106.

16. Quinlan JR. C4.5: Programs for machine learning. Morgan Kaufmann; 1993: 205.

17. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Wadsworth International Group; 1984: 124.

18. Martsenyuk VP, Semenets AV. Medical Informatics. Developer and Expert Systems, Ternopil, Ukrmedknyha, 2004: 222.

Received: 2014.05.07