Data Mining-a useful tool for fast and better clinical decision-making

A. Iordache, C. Surcel, R. Stoica, I. Sinescu
Center for Uronephrology and Renal Transplantation, Fundeni Clinical Institute, Bucharest

Information technology development has grown rapidly in the last years. The ***Data Mining (DM)*** technique has become an established method for improving statistical tools to predict future trends. The wide range of applications from business tasks to scientific tasks has lead to a huge variety of learning methods and algorithms for rule extraction and prediction. The general tasks of classification, regression, clustering, or deviation analysis have a large number of solutions such as neural networks, decision tree learners, rule learners or Bayesian networks.

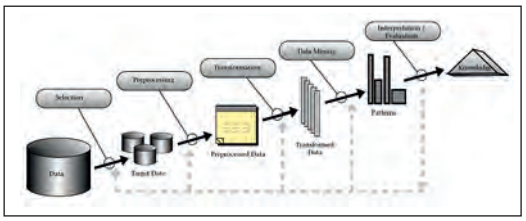
For medical diagnosis there are many expert systems based on logical rules for decision making and prediction. The most precise rules are often with low completeness as they are special cases and do not bring any help in prediction and future trends. The real medical applications are systems with imprecision and uncertainty in logical rules. Sophisticated equipment used in the practice of modern medicine generates huge amount of data. This data is usually stored in digital form, and considerable effort is being made to find automated methods of data analysis to generate knowledge which can be used for fast and better clinical decision-making.

Correspondence:

The area of data mining and knowledge discovery tools which is a well-defined process consisting of several steps including data mining can play an effective role in achieving these goals. The healthcare environment is usually information rich but knowledge poor. However, data mining techniques can be applied to create a knowledge rich healthcare environment. For example, applications of data mining techniques to Acquired Immune Deficiency Syndrome (AIDS) datasets can be a highly important area. In a thickly populated country with scarce resources such as India, information dissemination and knowledge discovery from large databases seem to be the only solution to check the spread of AIDS. In view of the large amount of medical data being generated, there is growing pressure for improved methods of data analysis and knowledge discovery using appropriate data mining techniques. A proper medical database created with intention mining can provide a useful resource for data mining and knowledge discovery.

The terms ***knowledge discovery in databases (KDD***) and **data mining** are used interchangeably, KDD being the process of finding useful information, and data mining, the process automatic or semiauto-

matic for extracting knowledge (information and patterns) derived by the KDD process.



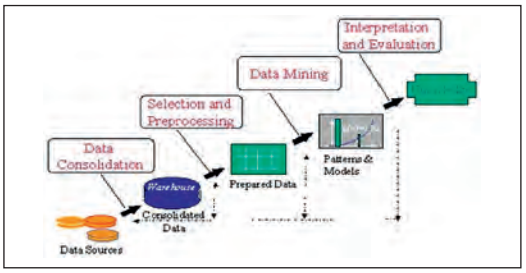
Beyond the problem of „pattern finding" and „machine learning" through algorithms

Medical data include data about a patient's history, laboratory reports, in-patient data, out-patient data, nursing staff information, specialists' information, other employees' records, surgery schedule, pharmacy information, billing, budget, etc. The computerized medical records, containing information about a patient's history, physical examination details, and laboratory findings are more legible and can be made available at several points in the hospital, than the traditional medical records paper. Medical data are different from data in other databases; they are heterogeneous and involve text and images to a great extent. For example, MRI, ECG or EEG, and cardiac SPECT generate huge amounts of heterogeneous medical data. Knowledge discovery from this data can greatly benefit mankind in the form of improved diagnostic techniques. The records of millions of patients can be stored and computerized. However, there are other important issues such as ownership and privacy related to these records. There is a need to develop methods for mining different types of these data including X-ray, MRI images, electrocardiogram ECG signals, and cholesterol level. The medical history of a patient is an important factor in determining the nature of treatment for the patient.

The large growth of medical databases in advanced countries has motivated medical researchers to use data mining for knowledge discovery from these databases. As the volume of stored data increases, data mining techniques assume an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Data mining techniques can help answer several critical questions, such as:

- ❑ given the records of dialysis patients, what can be done to improve the treatment of these patients?
- ❑ given the historical patient records on cancer, should the treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?
- ❑ could human DNA databases be characterized as genetic coding models ?

Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Without data mining, it is difficult to realize the full potential of data collected within an organization, as the data is massive, highly dimensional, distributed, and uncertain. In many medical diagnostics, data mining techniques can be used for knowledge discovery. Potential uses of data mining techniques, including artificial neural networks for medical diagnostics, have been effectively demonstrated by *(Scales & Embrechts, 2002) and (Kraft, Desouza, & Androwich, 2003).*



Data mining techniques form a group of heterogeneous tools and techniques and are used for different purposes. These techniques and methods are based on statistical techniques, visualization, machine learning, etc.

Simple apply ***Structured Query Language (SQL)*** to a dataset it could be obtain much statistical information such as:

- ❑ What is the average age of persons suffering from diabetes?
- ❑ What is the average age of persons suffering from heart disease?
- ❑ What is the average period of survival after angioplasty/ heart operation/ cancer treatment?
- ❑ What is the average age of persons admitted to a hospital suffering from a heart problem / cancer/ diabetes?
- ❑ What is the average hemoglobin of women of a particular community?

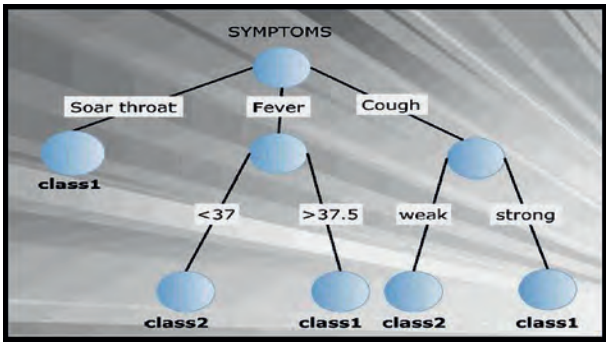
Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. Basically, the algorithms try to fit a model closest to the characteristics of data under consideration. Models can be predictive or descriptive.

Predictive models are used to make predictions,

for example, to make a diagnosis of a particular disease. A patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. **Descriptive models** are used to identify patterns in data. Classification, regression, and time series analysis are some of the tasks of predictive modeling, whereas clustering, association rules, and visualization are some are the tasks of descriptive modeling.

Usually, classification is a preliminary data analysis step for examining a set of cases to see if they can be grouped based on "similarity" to each other. Data analysis methods vary on the way how they detect patterns. The ultimate reason for doing classification is to increase understanding of the domain or to improve predictions compared to unclassified data. Given a classification and a partial observation, one can always use the classification to make a statistical estimate of the unobserved attribute values and as the departure point for constructing new models, based on user's domain knowledge.

The *decision-tree method* (used in systems CART, Clementine) like the ***nearest neighbor*** method, exploits clustering regularities for construction of decision-tree representation. It shows which variables are more significant with respect to classification decisions. The decision tree learning method requires the data to be expressed in the form of classified examples.



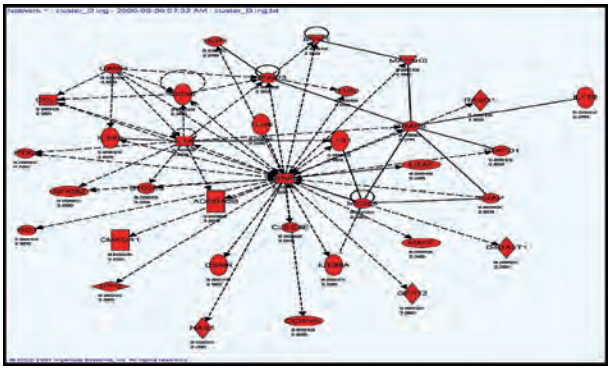
Decision tree (www. decision tree applet.com)

The method of ***Memory Based Reasoning*** also called the ***nearest neighbor method*** finds the closest past analogs of the present situation and chooses the same solution. Such systems demonstrate good results in vastly diverse problems. They do not create any models or rules summarizing the previous experience.

- ***Genetic Algorithms*** are powerful technique for solution of various combinatorial or optimization problems. They are more an instrument for scientific research, rather than a tool for generic practical data analysis.
- ***Nonlinear Regression Methods*** **(NR)** are based on searching for a dependence of the target variable in the form of function. Such methods provide solutions with a larger statistical significance than neural networks do. This method has better chances of providing reliable solutions in medical diagnostics applications.
- ***Support Vector Machines*** methods are based on Structural Risk minimization principle from statistical learning theory. They generalize better than NN.
- ***Maximum Likelihood Estimation*** **(*MLE*)** deals with finding the set of models and parameters that maximises this probability.

With statistical classification the usual procedure is to select the class, that most probably generates the partial observation and then use that class' distributions for the unknowns.

The improvement of new technologies rises data collection and accumulation, but without appropriate processing and interpretation this information remains useless. There are four standard methods for Data Mining: association, classification, clustering techniques and prediction.



A cluster network

For most medical applications the logical rules are not precise, and the uncertainty is present both in premise and in the decision. For this kind of application a good methodology is the rule representation from decision-tree method, which is easily understood by the user. The experiments show that induced decision trees are useful for the analysis of the importance of clinical parameters and their combinations for the prediction of the diagnosis. Clinical data when collected contains many errors and therefore needs to be standardized and checked for accuracy and reliability.

Computers are no doubt faster than humans in performing mathematical calculations, but human brains can perform many complex tasks such as image and speech recognition in a better way. For example, the artificial neural network, an important data mining technique, attempts to capture, to some extent, this aspect of brainpower in computer models.

Medical records are highly sensitive and contain a large amount of personal information. No doubt, traditional record keeping methods are also vulnerable regarding privacy and security, but remote access to a patient's data raises further risks of unauthorized entry into the data system. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis and prognosis. These data need to be collected in an organized form, which can be then integrated to form a hospital information system. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. Data mining and statistics both strive towards discovering patterns and structures in data.

Acknowledgement

This paper is supported by the Sectorial Operational Programme Human Resources Development (SOP HRD) 2007-2013, financed from the European Social Fund and by the Romanian Government under the contract number POSDRU/107/1.5/S/82839.

References

1. Ichise, R. & Numao, M. (2005) *First-Order Rule Mining by Using Graphs Created from Temporal Medical Data.* LNAI, Vol. 3430, pp. 112-25;
2. Kobayashi, T. & Kawakubo T. (1994) *Prospective Investigation of Tumor Markers and Risk Assessment in Early Cancer Screening. Cancer,* Vol. 73, No. 7, pp. 1946-1953;
3. Ohsawa, Y., Okazaki, N., & Matsumura, N. (2003) *A Scenario Development on Hepatics B and C. Technical Report of JSAI,* SIG-KBS-A301, pp. 177-182;
4. M. Ulieru, P. Palensky, and R. Doursat (Eds.): *IT Revolutions 2008,* LNICST 11, pp. 216–224, 2009;
5. Cios, K. J., & Moore, G. W. (2000) *Medical Data Mining and Knowledge Discovery: An Overview.* In Cios K. J., Medical Data Mining and Knowledge Discovery. Heidelberg
6. Scales, R., & Embrechts, M. (2002) *Computational Intelligence Techniques for Medical Diagnostics. Proceedings of Walter Lincoln Hawkins,* Graduate Research Conference 2012;
7. Lundin, J. (1998) *Artificial Neural Networks in outcome prediction.* Anns Chir Gynaecol 87, 128-130.;
8. Itchhaporia, D., Snow P.B., Almassy, R. J., & Oetgen, W. J. (1996) *Artificial Neural Networks: current status in cardiovascular medicine.* J Am Coll Cardiol. Aug, 28(2), 515-521;