Data mining in healthcare: decision making and precision

Ionuț ȚĂRANU
University of Economic Studies, Bucharest, Romania
ionut.taranu@gmail.com

The trend of application of data mining in healthcare today is increased because the health sector is rich with information and data mining has become a necessity. Healthcare organizations generate and collect large volumes of information to a daily basis. Use of information technology enables automation of data mining and knowledge that help bring some interesting patterns which means eliminating manual tasks and easy data extraction directly from electronic records, electronic transfer system that will secure medical records, save lives and reduce the cost of medical services as well as enabling early detection of infectious diseases on the basis of advanced data collection. Data mining can enable healthcare organizations to anticipate trends in the patient's medical condition and behaviour proved by analysis of prospects different and by making connections between seemingly unrelated information. The raw data from healthcare organizations are voluminous and heterogeneous. It needs to be collected and stored in organized form and their integration allows the formation unite medical information system. Data mining in health offers unlimited possibilities for analyzing different data models less visible or hidden to common analysis techniques. These patterns can be used by healthcare practitioners to make forecasts, put diagnoses, and set treatments for patients in healthcare organizations.
***Keywords**: Data Mining, Big Data, Knowledge Discovery*

1 Introduction

Health organizations today are capable of generating and collecting a large amount of data. This increase in data volume automatically requires the data to be retrieved when needed. With the use of data mining techniques is possible to extract the knowledge and determine interesting and useful patterns. The knowledge gained in this way can be used in the proper order to improve work efficiency and enhance the quality of decision making. Above the foregoing is a great need for new generation of theories and computational tools to help people with extracting useful information from the growing volume of digital data [1]. Information technologies are implemented increasingly often in healthcare organizations to meet the needs of physicians in their daily decision making. Computer systems used in data mining can be very useful to control human limitations such as subjectivity and error due to fatigue and to provide guidance to decision-making processes [2]. The essence of data mining is to identify relationships, patterns and models which support predictions and decision-making process for diagnosis and treatment planning. These can be called predictive models, and integrated in hospitals information systems as models of decision making, reduce subjectivity and the necessity for reducing the time for decision making. In addition, the use of information technology in healthcare enables the comprehensive management of medical knowledge and its secure exchange between healthcare providers and beneficiaries [3]. Obtaining information using computers can help the quality of decision-making and avoiding human error. When there is a large volume of data that must be processed by the people, making decisions is generally of poor quality [4]. Data mining is the process of analyzing the raw data using a computer and extracts their meaning. The process is often defined as the discovery of previously unknown and potentially useful information from large volumes of data (unstructured) [5].

Thanks to this technique, it is possible to predict trends and behavior of patients or diseases. This is done by analyzing data from different perspectives and finding connections and relationships between seemingly unrelated information. In the process of data mining previously unknown trends and patterns from a database of information are discovered and transform information into meaningful solutions [6].

2. Data mining and Knowledge discovery process

Knowledge Discovery (KDD) is a process that allows automatic scanning of high-volume data in order to find useful patterns that can be considered knowledge about the data (Fig 1)**.** Once discovered knowledge are presented, evaluation methods can be improved, data mining process can be further "refined", new data can be selected or subsequently processed, and new data sources can be integrated in order to get different results corresponding to [7]. This is the process of converting low level information into knowledge of high level. Therefore, KDD is a non-trivial extraction of implicit information, previously unknown and potentially useful data is in the database. Although data mining and KDD are often treated as equivalent, in essence, data mining is an important step in the KDD process. Knowledge discovery process involves the use of the database, along with any selection, pre-processing, sub-sampling and transformation; application of data mining methods to enumerate the models; evaluation of the data mining product to identify subsets listed models representing knowledge. Data mining component knowledge discovery process refers to algorithmic means by which patterns are extracted and listed from the available data [1].

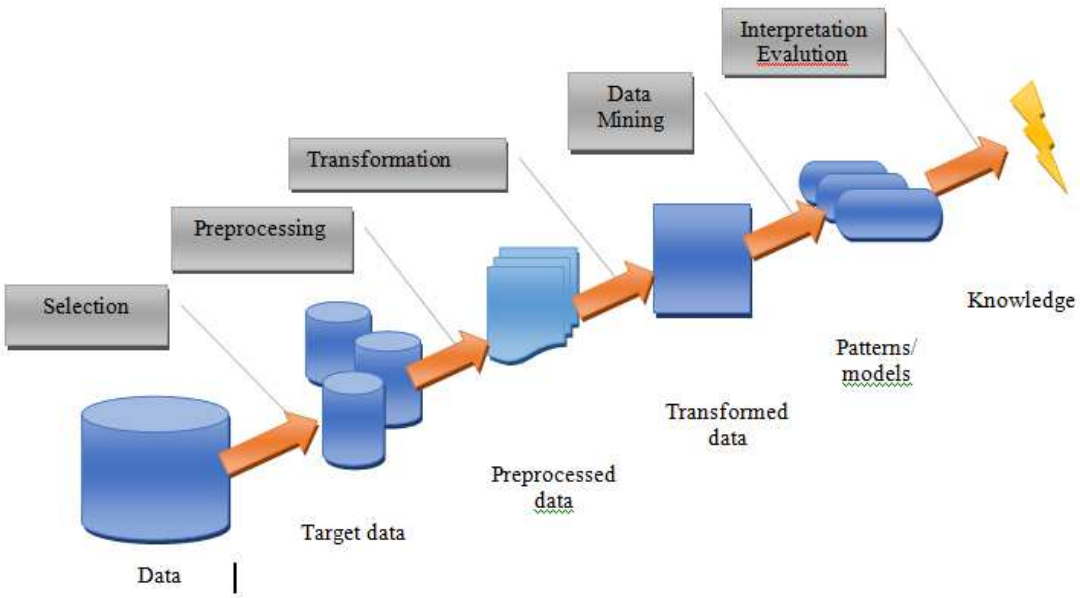

Fig. 1 Knowledge discovery process.

The daily amount of information that it stores by large organizations in their databases is measured in terabytes. 1 terabyte can store text equivalent to approximately two million books. However, these raw data, poorly structured, with different formats, is not very useful. It is necessary that the data is processed and analysed, and based on these processing actions following to extract useful information [5]. There have been developed a number of models and algorithms for autonomous prediction based on data corresponding to various features [8].

Different methods serve different purposes, each method having its advantages and

disadvantages. Data mining tasks can be divided into descriptive and predictive [9]. While descriptive tasks aim to find a human interpretation of forms and associations, after reviewing the data and the entire construction of the model, prediction tasks tend to predict an outcome of interest. The main tasks of predictive and descriptive data mining can be classified as follows [10]:

• Classification and Regression - identification of new templates with predefined objectives; These tasks are predictive and they include the creation of models to predict target, or dependent variable from the set of explained or independent variables.

• Association rule – association rule analysis type represents a descriptive task which includes determining patterns, or associations, between elements in data sets

• Cluster analysis – descriptive data mining task with the goal to group similar objects in the same cluster and different ones in the different clusters.

• Text mining – most of the available data is in the form of unstructured or partially structured text, and it is different from conventional data that are completely structured. While text mining tasks usually fall under classification, clustering and association rule data mining categories, it is the best to observe them separately, because unstructured text demands a specific consideration. In particular, method for representation of textual data is critical.

• Link analysis – Form of network analysis that examines the associations between objects. Link classification provides category of an object, not just based on its features, but also on connections in which it takes part, and features of objects connected with certain path [11].

3. Application of data mining in healthcare

Healthcare abounds various information which causes the necessity of data mining application. It is well known that healthcare is a complex area where new knowledge is being accumulated daily in a growing rate. Big part of this knowledge is in the form of paperwork, resulting from a studies conducted on data and information collected from the patient's healthcare records. There is a big tendency today to make this information available in electronic form, converting information to knowledge, which is not an easy thing to do [12]. All healthcare institutions need an expert analysis of their medical data, project that is time consuming and expensive [13]. The ability to use a data in databases in order to extract useful information for quality health care is a key of success of healthcare institutions [4]. In medical research, data mining begins with the hypothesis and results are adjusted accordingly, different from standard data mining practice, that begins with a set of data without obvious hypothesis [14]. While the traditional data mining is focused on patterns and trends in data sets, data mining in healthcare is more focused on minority that is not in accordance with patterns and trends. The fact that standard data mining is more focused on describing and not explaining the patterns and trends, is the one thing that deepens the difference between standard and healthcare data mining. Healthcare needs these explanations since the small difference can stand between life and death of a patient.

Here are some of the techniques of data mining, which are successfully used in healthcare, such as artificial neural networks, decision trees, and genetic algorithms and nearest neighbour method.

Artificial neural networks are analytical techniques that are formed on the basis of superior learning processes in the human brain. Neural networks are groups of connected input/output units where each connection has its own weight (*Fig 2*) [15].

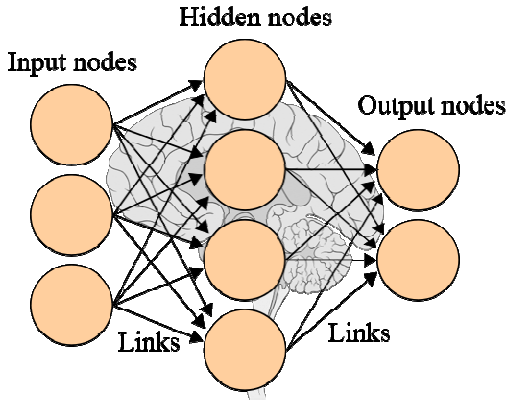

Fig. 2 Artificial neural network

The learning process is performed by balancing the net on the basis of relations that exist between elements in the examples. Based on the importance of cause and effect between certain data, stronger or weaker connections between "neurons" are being formed. Network formed in this manner is ready for the unknown data and it will react based on previously acquired knowledge.

Decision tree (Fig 3) is a graphical representation of the relations that exist between the data in the database. It is used for data classification. The result is displayed as a tree, hence the name of this technique. Decision trees are mainly used in the classification and prediction. The instances are classified by sorting them down the tree from the root node to some leaf node [16]. The nodes are branching based on if-then condition. Tree view is a clear and easy to understand, decision tree algorithms are significantly faster than neural networks and their learning is of shorter duration. Decision tree can also be interpreted as a special form of a rule set, which is characterized by its hierarchical organization of rules.

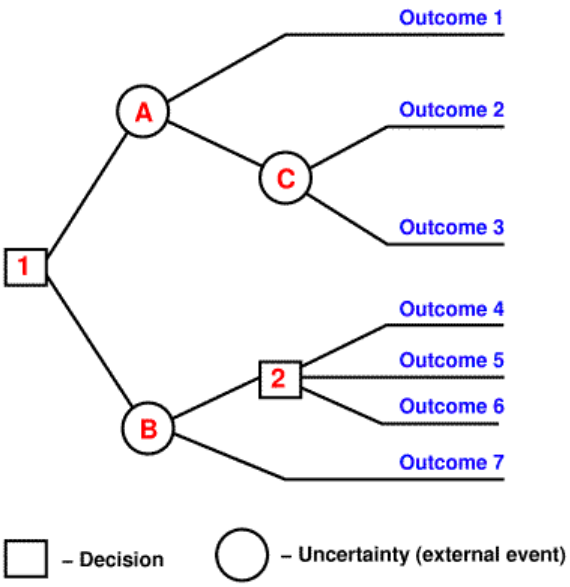

Fig. 3 Decision tree analysis

Genetic algorithms are based on the principle of genetic modification, mutation and natural selection. These are algorithmic optimization strategies inspired by the principles observed in natural evolution [15]. Genetic algorithms are used in data mining to formulate hypotheses about the dependencies between variables in the form of association rules or other internal formalism [17].

Nearest neighbour method (Fig 4) is a technique that is also used for data classification. Unlike other techniques, there is no learning process to create a model. The data used for learning is in fact a model. When the new data shows up, the algorithm analyses it to find a subset of instances that are the best fit and based on that it is able to predict the outcome. The study [18] conducted on the application of nearest neighbour method on benchmark data set to detect efficiency in the diagnosis of heart diseases, revealed that application of this method had an accuracy of 97.4% which is a higher percentage than any other published study on the same set of data.

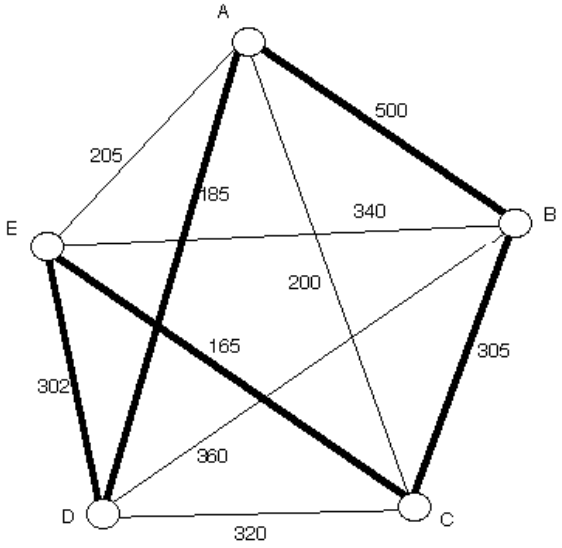

Fig. 4 Nearest Neighbour Algorithm

3.1 Advantages

Information system simplifies and automates the workflow of health care institution.

Integration of data mining in information systems, healthcare institutions reduce subjectivity in decision-making and provide a new useful medical knowledge. Predictive models provide the best knowledge support and experience to healthcare workers. The goal of predictive data mining in medicine is to develop a predictive model that is clear, gives reliable predictions, support doctors to improve their prognosis, diagnosis and treatment planning procedures.

A very important application of data mining is for biomedical signal processing expressed by internal regulations and responses to the stimulus conditions, whenever there is a lack of detailed knowledge about the interactions between different subsystems, and when the standard analysis techniques are ineffective, as it is often the case with non-linear associations [19].

3.2 Obstacles

One of the biggest problems in data mining in medicine is that the raw medical data is voluminous and heterogeneous [20]. These data can be gathered from various sources such as from conversations with patients, laboratory results, review and interpretation of doctors. All these components can have a major impact on diagnosis, prognosis and treatment of the patient, and should not be ignored. Missing, incorrect, inconsistent or non-standard data such as pieces of information

saved in different formats from different data sources create a major obstacle to successful data mining.

Also, another obstacle is that almost all diagnoses and treatments in medicine are imprecise and subjected to error rates. Here the analysis of specificity and sensitivity are being used for the measurement of these errors. Within the issue of knowledge integrity assessment, two biggest challenges are: (1) How to develop efficient algorithms for comparing content of two knowledge versions (before and after). This challenge demands development of efficient algorithms and data structures for evaluation of knowledge integrity in the data set; and (2) How to develop algorithms for evaluating the influence of particular data modifications on statistical importance of individual patterns that are collected with the help of common classes of data mining algorithm. Algorithms that measure the influence that modifications of data values have on discovered statistical importance of patterns are being developed, although it would be impossible to develop a universal measure for all data mining algorithms [21].

4. Conclusions

Data mining has great importance for area of medicine, and it represents comprehensive process that demands thorough understanding of needs of the healthcare organizations.

Knowledge gained with the use of techniques of data mining can be used to make successful decisions that will improve success of healthcare organization and health of the patients. Data mining requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results. Data mining, once started, represents continuous cycle of knowledge discovery. For organizations, it presents one of the key things that help create a good business strategy. Today, there has been many efforts with the goal of successful application of data mining in the healthcare institutions. Primary potential of this technique lies in the possibility for research of hidden patterns in data sets in healthcare domain. These patterns can be used for clinical diagnosis. However, available raw medical data are widely distributed, different and voluminous by nature. These data must be collected and stored in data warehouses in organized forms, and they can be integrated in order to form hospital information system. Data mining technology provides customer oriented approach towards new and hidden patterns in data, from which the knowledge is being generated, the knowledge that can help in providing of medical and other services to the patients. Healthcare institutions that use data mining applications have the possibility to predict future requests, needs, desires, and conditions of the patients and to make adequate and optimal decisions about their treatments. With the future development of information communication technologies, data mining will achieve its full potential in the discovery of knowledge hidden in the medical data.

References

[1] Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence , 37-54.

[2] Candelieri, A., Dolce, G., Riganello, F., & Sannita, W. G. (2011). Data Mining in Neurology. In Knowledge-Oriented Applications in Data Mining (pp. 261-276). InTech.

[3] Bushinak, H., AbdelGaber, S., & AlSharif, F. K. (2011). Recognizing The Electronic Medical Record Data From Unstructured Medical Data Using Visual Text Mining TechniquesProf. Hussain Bushinak. (IJCSIS) International Journal of

Computer Science and Information Security, Vol. 9, No. 6 , 25-35.

[4] Eapen, A. G. (2004). Application of Data mining in Medical Applications. Ontario, Canada, 2004: University of Waterloo.

[5] Milovic, B. (2011). Usage of Data Mining in Making Business Decision. YU Info 2012 & ICIST 2012, (pp. 153-157).

[6] boirefillergroup.com. (2010). Data Mining Methodology. Retrieved 06 12, 2012,from Boire-Filler Group: http://www.boirefillergroup.com/methodology.php

[7] Zaïane, O. R. (1999). Principles of Knowledge Discovery in Databases. Department of Computing Science, University of Alberta.

[8] Kusiak, A., Kernstine, K., Kern, J., McLaughlin, K., & Tseng, T. (2000). Data Mining: Medical and Engineering Case Studies. Industrial Engineering Research 2000 Conference, (pp. 1-7). Cleveland, Ohio.

[9] Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. international journal of medical informatics 77 , 81–97.

[10] Weiss, G. M., & Davison, B. D. (2010). Data Mining. Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons .

[11] Getoor, L. (2003). Link Mining: A New Data Mining Challenge. SIGKDD Explorations Volume 4, Issue 2 .

[12] Ceusters, W. (2001). Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare. KJ Cios (ed.) Medical Data Mining and Knowledge Discovery, Physica-verlag Heidelberg, (pp. 41-67). New York.

[13] Matheus, C. J., Shapiro, G. P., & McNeill, D. (1996). Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press , 445-463.

[14] Canlas, R. D. (2009). Data Mining in Healthcare: Current Applications and Issues. Carnegie Mellon University, Australia.

[15] Gupta, S., Kumar, D., & Sharma, A. (2011). Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis. Indian Journal of Computer Science and Engineering (IJCSE) 188-195.

[16] Khan, F. S., Anwer, R. M., Torgersson, O., & Falkman, G. (2008). Data Mining in Oral Medicine Using Decision Trees. World Academy of Science, Engineering and Technology 37, (pp. 225-230).

[17] Ngan, P. S., Wong, M. L., Lam, W., Leung, K. S., & Cheng, J. C. (1999). Medical data mining using evolutionary computation. Artificial Intelligence in Medicine 16, (pp. 73–96).

[18] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. 2012 International Conference on Knowledge Discovery (ICKD 2012) IPCSIT Vol. XX. Singapore: IACSIT Press.

[19] Stühlinger, W., Hogl, O., Stoyan, H., & Müller, M. (2000). Intelligent Data Mining for Medical Quality Management. Proc. Fifth Workshop Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP2000), Workshop Notes of the 14th European Conf. Artificial Intelligence.

[20] Cios, K. J., & Moore, G. W. (2002). Uniqueness of Medical Data Mining. To appear in Artificial Intelligence in Medicine journal .

[21] Yang, Q., & Wu, X. (2006). 10 Challenging problems in data mining research. International Journal of Information Technology & Decision Making Vol. 5, No. 4 , 597–604.

Ionuţ ŢĂRANU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1996, having its Master degree on “Database support for business”. At present is in the process of getting his title of doctor in economy in the specialty of “Soft-computing methods for early medical diagnosis”. He has been an Assistant Professor for 4 years at “Titu Maiorescu” University and also for 4 years at Academy of Economic Studies from Bucharest. He published a series of articles, from which the most important are Applying ABCD Rule of Dermatoscopy using cognitive systems and ABCDE Rule in Dermoscopy – Registration and determining the impact of parameter E for evolution in diagnosing skin cancer using soft computing algorithms.

Mr. Taranu is currently the General Manager of Stima Soft company. He has more than 15 years of experience as a project manager and a business analyst with over 13 years of expertise in Software development, Business Process Management, Enterprise Architecture design and Outsourcing services. He is also involved in research projects, from which the most relevant are:

- Development of an Intelligent System for predicting, analyzing and monitoring performance indicators of technological and business processes in renewable energy area;
- Development of an eHealth platform for improving quality of life and the personalization of therapy at patients with diabetes;
- Development of an Educational Portal and integrated electronic system of education at the University of Medicine and Pharmacy "Carol Davila” to develop medical performance in dermatological oncology field;