Review Article

Data Mining, an Approach for Developing the Health Domain

Shahram Tofighi[1], Ali Ghazvini[2], Gholamhossein Pourtaghi[3], Mansour Esmaeilpour[4], Reza Shahhoseini*[5]

Abstract

Nowadays, data mining as the process of arrangement and classifications of voluminous data is one of the most important technics for studying and analyzing data in different organizations and domains. Data mining is among technological improvements towards data managing. Also, the wide use of information systems and databases has converted its merging with traditional methods into a necessity. Due to the existence of the large datasets in health-care organizations, data mining process has become necessary towards the automatic summarization of data and the extraction of the stored information and detection of the pattern from data. As nowadays the wide volume of data is daily obtained during care and treatment processes, analyzing them in order to discover the patterns and new science that can be resulted to upgrade health has been extremely inconspicuous. Therefore, the purpose of the present research is studying strategies and technics of data mining as one of the most important approaches in the development of health domains.

Keywords: Data Mining, Health, Decision Tree, Neural Networks, Networks, Genetic Algorithms

1. Health Management Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran
2. Medical School, Baqiyatallah University of Medical Sciences, Tehran, Iran.
3. Health Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran.
4. Computer Engineering Department, College of Engineering, Hamedan Branch, Islamic Azad University, Hamedan, Iran.
5. Health Care Management, Faculty of Health Baqiyatallah University of Medical Science, Tehran, Iran.

* **Corresponding Author**
Reza Shahhoseini, Health Care Management, Faculty of Health Baqiyatallah University of Medical Science, Tehran, Iran
E-mail: reza_shahhoseini@yahoo.com

Submission Date: 10/28/2014
Accepted Date: 12/23/2014

Introduction

During the past two decades, human's technical power has quickly increased to generate and collect data. Factors such as the wide use of barcodes for commercial products, possessing computer into business, sciences, governmental services and improvement of data collection instruments, from scanning texts and images to the satellite remote sensing systems, have an important role in these changes (1). In general, the public use of Web and Internet as a global informing system counters us to the large volume of data and information. This explosive growth in the stored data has made urgent requirements to the existence of modern technologies and automatic tools that intelligently help human beings to transform this large volume of data to information and knowledge (2). Data mining is posed as a solution for these problems. In an informal definition, data mining is said to be an automatic process for the extraction of patterns which represent knowledge. This knowledge has implicitly been stored in large databases, data stores, and other large resources of information.

Data mining instantaneously benefits from some scientific fields such as: database technology, artificial intelligence, machine learning, neural networks, statistics, pattern identifying, knowledge-based systems, knowledge reaching, information retrieval, high speed computations and visual representations of data (3). Data mining has emerged at the late 1980s, the important steps have been surveyed in this scientific branch in the 1990s and it expected to continue its growth and improvement in this century. In the information explosion era, the private companies will daily produce and collect large volumes of data. The extraction of useful information from databases and transforming information to practical results is a main challenge which companies encounter these days. Regarding the countries' improvements in the context of information technology, special insights towards an electronic government and the influence of computer systems in industries and the creation of large informational banks by governmental offices and privative departments, have all caused data mining to become a necessary requirement.

The health industry is continuously generating scales of data (4), and the people who encounter these types of data, have understood that there is a wide gap between data collection and the interpretation of them (5).The rather young and growing domain of data mining is among methods which can impart this industry from profound analysis of these data and result to the development of medical researches and scientific decisions in the context of diagnosis and treatment (6).

Concept of data mining

Since the appearance of statistical science, the scientists have sensed the need to discover data properties. As the need to use data and the perception of information has increased, the data have been quickly collected and stored with an increasing speed (7). The large volume of stored data and the extension of its dimensions have had different formats that have caused many problems and the statistical methods aren't singly able to discover the data properties. For solving this problem, the scientists decided to use the high speed of computers and other statistical methods such as the neural networks and genetic algorithms. The three themes of data storing, the increment of the computers speed and the genesis of the modern algorithms have caused to create a science namely data mining (8). Data mining can counted arising from natural evolutional trend of information technology which this evolutional trend is arising from an evolutional trend in the database industry, such as the operations: data collection and database making, data management and analysis and understanding of data (9).

Data mining is the process of the automatic or semi-automatic analysis of large quantities of data in order to discover the meaningful patterns and rules (10). Data mining is the process of the extraction and recognition of the hidden patterns or information from databases. In a better expression, the mechanized analysis of data in order to find useful, fresh and documentable patterns in large databases is named data mining (11). Data mining can be defined as the process of employing a computer-based methodology that directly extracts knowledge from data by using the different technics. Data mining creates technologies as data storage and manages software for the management in order to gain competitive advantages (12).

The goal of data mining is to discover valid, new and detectable ideas in the numerous volume of data by using statistical tools and artificial intelligence (13). The history of knowledge discovery in the information bases that is famous as data mining, doesn't have antiquity. In the early 1990s, when the knowledge discovery term was proposed in the informational bases for the first time, the companies had proceeded to store large quantities of data and the public invasion formed towards designing the data mining algorithms (12). It was in this time when the concept of data mining and its tools were concentered and followed the methods for the productivity of data stores. The main mission of data mining is accomplishing two total classes: description and prediction. In the description level, the goal is finding data of past and present times.

The description patterns are utilized to search a group of the similar variables among people or sets of common demographic groups. Prediction is used for assessing about the unknown affairs on the basis of the known affairs too. This specification can be used for future predictions or the assessment about the present. There are two functions in prediction: classification: that its purpose is inserting an item in a class; and estimation: that its purpose is to produce numerical quantities for an unknown variable (13). Therefore, it observed that everybody based on the application and usage cases, has provided a definition of data mining. Of course, since many years ago, statisticians have typically used data mining in different names such as data fishing, data dredging and data studying. Despite data mining is a modern scientific field, but nowadays it has acquired various and wide applications in the fields such as marketing, medicine, engineering, computer sciences, industry, quality control, communication and agriculture.

Main parts of the data mining system

Nowadays, the medical data volume which are stored electronically are increasing. Unfortunately large sets of raw data have no applications in itself. The most basic reason which has made data mining a focused subject in medical sciences, is the problem of the availability of vast volumes of data and the serious need to extract useful information and knowledge from these data. The data mining system has many different parts (14) such as:

Database, data store and other information resources: It performs through a set of databases, data stores, wide planes, and other types of information resources, data purging, and integration technics.

Servicer of the database or data store: Which is responsible of the related data according to the type of the user's data mining request.

Knowledge base: This base has consisted of the context knowledge to help search or it is used for the evaluation of the found patterns.

Data mining motor: This motor is the main part of the data mining system and it ideally includes the measures such as description, association, classification, analysis of the clusters, and analysis of the evolution and deviation.

Measure of pattern evaluating: This part applies the attraction criterions and interacts with the data mining measure, so that its concentration is on searching among the attractive patterns and utilizes a threshold limit for the assessment of the discovered patterns.

User's graphical interface: This measure makes a relation between the user and the data mining system, allows users to connect with the data mining system through query. This part allows users to review the visage of database or data store, evaluate the found patterns and represent the patterns in various visual forms (15).

The necessity and importance of data mining

The human power for the perception and understanding of data, the quick and salient growth of data, collecting and storing them in numerous databases aren’t possible without strong tools. The collected data in databases have been converted into data sepulcher. As a result, important decisions weren't based on the rich stored information in the databases and the deciders didn’t have tools for extracting the knowledge hidden in large databases (16).

Nowadays, the scale of the available data becomes twice every 5 years and organizations are considered capable which can manage lower than 7 percent of its information (17). The major reason that data mining has been concentrated on by the information industry in recent years, is that it deals with a huge volume of data in a wide scale. This is while deciders are unable to collect rich information, impart resources for making up the actual decisions and despite the availability of the commercial data, survive in the lack of commercial knowledge (18).

It is obvious and clear that vast volume of data has been collected, however it is questionable that what is learned from this data? What knowledge is gained from this information? In the beginning of 1984, John Naisbitt stated in the Megatrends Book that "we are drowning in information, but starved for knowledge". In fact, we are full of data in most of the fields and the problem is that we don't possess enough analyzers which require skill and experiment for transforming data to knowledge (19). The considerable growth in the contexts of data mining and knowledge extracting has been invigorated from the junction of various factors and causes:

- Progressive growth of data collection; data are going towards being generated and available in large volumes
- Storing data and the development of the analytic online processing technology
- Increment of the access scale to the data of Web and Internet
- The salient growth of the computation power and space of information storing
- Development and deployment of software products in data mining. Also, the commercial softwares of data mining are easily accessible and the applied programs have been as an interface for the standard users of data mining.
- Incrementing competition tensions in the stock markets of the world economy and interesting to customer-holding management
- In the vast range of industries, the commercial companies and institutions have found out that the customers are the major core of marketing and customers' information is considered as one of their investment keys (18).

Many organizations whether in private or governmental departments, have been attracted by data mining methods by using technologies and processes of the progressive business. Some of these changes consist of: growth of computer networks, connection to databases, and development of search technics such as neural networks and advanced algorithms, deployment of the customer-worker models, increasing users' access scale to data central resources and increment of the ability of data compounding from various resources and transforming it to a unity searchable resource (20).

Furthermore, the organizations use data mining as a tool in customer, economy and medical researches. Data mining consists of two main steps: data preprocessing and pattern recognition. Data processing includes the cases that compounds the number of data elements are much and or the marks are derivation of several simple data. Usually the preprocessing process is time-consumer (21) and pattern recognition is used where the data pattern is compound. Regarding to the quick increment of the computers power in the past decades and the increment of the number of large datasets and recognizing the value of other smooth changes, the traditional methods aren't singly able to present the powerful analyses of data and this emphasizes the necessity of the computer analysis methodology (22).

Application of data mining in Different Fields

The base and fundament of data mining is rooted in three old branches which classic statistics is the most important of them. Data mining doesn't exist without statistics as statistics is the infrastructure of most technologies which data mining has been built on (23). In general, data mining is an interdisciplinary movement which encompasses the domains such as: the information bases, statistics, mechanized learning, quick computation, visualization, and mathematics (24). Specially, data mining is a branch of limited intelligence, but it has been successfully used for several years in commercial and scientific societies. It has been used for the detection of the groups and people's behavior, processing medical information, support services of customers, support of decision, and many other performances. For example, the data mining's principles have been led to special medical discoveries such as the relationship between estrogen and Alzheimer’s disease, and the relationship between migraine headaches and meningitis putrefaction (25).

In some departments such as commerce and medicine, in order to achieve commercial goals and also to predict the future outputs, data mining programs are used to extract useful information among voluminous data. In the medicine department, this affair is utilized as the use of the statistical tools for studying the drugs' effect on the diseases. The pharmacy companies use data mining of the chemical compounds and genetic materials to help research about the modern treatments for the diseases. The large commercial companies use the data mining tools to analysis the details of various months' transactions in order to infer the effect of price changes on goods profit in national levels (26). Also, these companies can design the models and predict costumers' behavior by using computer data which has been gathered during many years.

The companies such as the providers of telephonic services and music clubs can use data mining in order to recognize which customers will probably join and which customers will possibly pursuit to other competitors (27). The number of applied projects in the data mining area is increasing progressing. It is predicted that a type of sharp rise (about three hundred percent) occurred in the data mining projects towards an improvement of the customers' relations and support of customers' desires in the next decade (12). The researches have performed the use of data mining technics in order to detect the existent implicit knowledge in the news and media flows (28).

Data mining in health fields

What has been very important nowadays isn't the deficit or lack of required data but rather is deficit or lack of suitable or standard methods in order to preserve, update, pose available, and in a more ideal state, discovering modern knowledge from present data. The use of the data mining systems is one of the proposed strategies for reaching this goal. The data mining system allows users to interpret the collected data and extract the hidden knowledge (29). Although knowledge discovery entered health fields in order to recognize financial defalcation, but it was gradually used in clinical domain as well. This important subject is arising from the quick change of the consciousness relative to information in health fields (30).

Data mining in medicine and biology is an important section of bio-medicine informatics and computer science has been one of the most applied sciences which has been utilized in the hospitals, clinics, laboratories, and research centers (6). Data mining has slowly but progressively been employed to remove the multiple problems in knowledge discovery and health departments. The four most important reasons of the slow growth of this science in health fields are: the sensitivity of the science of medicine and its entwining with humans’ lives (The partial difference in the data mining patterns can be resulted to change the equilibrium between death and life), amazement in the definition of data mining (sometimes making a simple design of the medical databases is erroneously proposed as the pattern resulted of data mining),personal privacy and secrecy of health data and finally the most important challenge is that if is supposed the gaining data mining results which are perfectly trustworthy. It can be said that changing the habit of the care providers from traditional medicine to evidence-based medicine is difficult (31).

The data mining analyses are performable in two methods as with-supervisor and without-supervisor and through algorithms such as neural networks, classification and the genetic decision tree. In addition to these current algorithms, the new algorithms are generated for the scientific or commercial researches' goals through academic research projects. The exclusive properties of data mining can be recounted as follows (19):

- Not only affect the analysis phase, but also designing study and data collection.
- Provide the possibility of searching the answers for the accurate and high complicated questions in the collected data.
- Are obviously and clearly able to answer questions.
- Their main advantage and difference in compared to other technics is that instead of a mere presentation of the huge strategy, provides accurate answers for the researcher.
- Provides measuring possibilities of the different variables' effect on dependent variables.
- Helps managers evaluate the effect of future scenarios and proceeds to select the movement path by modeling various choices and also helps decision making in uncertainty conditions (29).

Decision tree: This technic has a structure similar to a tree that describes the rules which has resulted to the decision and description ease is of its important specifications. For example, the decision tree can specify the parameters effective on the survival scale of kidney grafts Also, the use of algorithm DRG in paying back the health insurance costs of elderlies in the United States is considered as a classical example for the methodology of this technic. Decision trees which are used to predict classic variables, are named classification trees, because they place the samples in classes or categories. The decision trees which are used to predict the continuous variables, are named regression trees (14).

Neural networks: This technic generates the nonlinear prediction models that instructs how a pattern is adaptable with a special profile, but it doesn’t provide any explanations about the causes of such special results. For example, neural networks can specify which type of diseases may be accompanied by other diseases (31) and helps diagnosis, treatments and drug manufacturing by analyzing the images, cardiograms and other clinical observations (32).The main purpose of this method is to find the set of weights for the network so that all the basic training to correctly classify or predict (33)

Fuzzy logic: The fuzzy logic is more flexible in compared to the other technics and manages the ambiguous and complicated concepts. These algorithms are obtained from the neural network's pattern. The main advantage of the fuzzy neural network is the ability in modeling special problems by using an oral high-level intelligible model, unlike the complicated mathematical phrases (14).

Genetic algorithms: These are optimum technics for upgrading other data mining algorithms so that the best model on datasets are used and can specify the best therapeutic schedule for a special disease. This algorithm is based on the Darwin's evolutionary hypothesis and its application is stabled on natural genetic.

The primary principles of the genetic algorithm has been presented by Holland et al in the year 1962. On this base, the more-compatible livings with environment, survive and reproduce at higher scale, consequently their presence chance is more in the next generations (15).

Application of data mining in health Area

The health industry is continuously generating large scales of data and the people who encounter with this type of data, have found out that there is a wide gap between the collection and the interpretation of them. Nowadays, health departments are most needy for data mining and the motion from traditional medicine towards evidence-based medicine is among the cases which can be emphatic for this affair. For the most important applications of data mining in the health domain, one can refer to the following:

Data mining in non-invasive diagnoses: Some diagnostic and laboratorial actions for patients are invasive, expensive and yet painful. For example, the biopsy of the cervix of uterus in order to diagnose cervical cancer is among these cases.

Data mining in the determination of the treatment type: Applying data mining on medical data has presented vital and effective consequences for selecting the suitable type of treatment and saving lives.

Data mining in the electronic file of health: Nowadays, multiple studies emphasis that the data mining technics provide effective tools for recognizing the important patterns of health among the medical files.

Data mining in hospitals ranking: The hospitals’ rankings and health programs can be based on the reported information by the care provider. Therefore, standard reporting is essential for meaningful comparisons of hospitals and their rankings. The use of data mining technics is among the standard methods (34). One can recognize the patients with dangerous conditions by using data mining and data modeling.

In fact, by providing information for care providers, data mining helps them recognize the unsafe patients and improve their care quality to prevent future problems. This will result in decreasing the hospitals’ acceptations by designing suitable interventions (35).

Discussion and Conclusion

Extracting information and knowledge from data is an old concept in scientific and medical researches and the new subject is the convergence and unity of many fields and corresponding technologies that has made an exclusive opportunity for data mining. Data mining has a vital role in health by making evidence-based medicine and is resulted in discovering the modern, helpful and stable knowledge in the databases of the health organizations. It is said that in order to achieve evidence-based medicine, one must commence from recognizing the gap and lack of the knowledge in the care processes of modern health and then follow the best reasons. In the next step, one must proceed to study the correctness and validation of the recognized actions and finally, perform these reasons on patients. Data mining grades the achievement way towards the first step in this context.

Regarding that at the moment vast volumes of daily data is gained during care and treatment processes in our country, but, the analysis and interpretation of them in order to discover new patterns and knowledge that can lead to upgrade health is very inconspicuous. Data mining has been designed as an analytic process for mining the designed data. Meanwhile searching the compatible patterns and systematic relations between the variables, can proceed to exploratory analyzing of the data, discovering the patterns and rules and algorithms, predictive modeling and searching the deviations.

References

1. Moqaddasi H, Hosseini A, Asadi F, Jahanbakhsh M. (1390), "Data mining and its application in health. Management of health information", Vol.9, No.2, [In Persian].

2. Daniel T. Larose, (2004). "Discovering Knowledge in Data: An Introduction to Data Mining".

3. Han, J and Kamber. M, (2001). "Data Mining: Concept and Techniques. Morgan Kaufmann".

4. Zhu L, Wu B, Cao C, (2003). Introduction to medical data mining. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi; 20(3): 559-62.

5. Cios KJ, (2000). "Medical data mining and knowledge discovery", IEEE Eng Med Biol Mag; 19(4): 15-6.

6. Berka P, Rauch J, Zighed DA. (2009)," Data Mining and Medical Knowledge Management: Cases and Applications", Hershey: Idea Group Inc (IGI).

7. Chen, YH & Su, CT. ' (2006), "A kano CKM model for customer knowledge discovery", Total Quality Management & Business Excellence, vol. 17, no.5, pp.589-608

8. Shu, H. Pei-H. Chu, Y. (2012), "Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications". doi:10.1016/j.eswa. 2012.02.063

9. Tan J. (2008), "Medical Informatics: Concepts, Methodologies, Tools, and Applications". Hershey: IGI Global snippet.

10. Wager KA, Lee FW, Glaser JP. (2005), "Managing Health-care Information Systems: A Practical Approach for Health-care Executives". New Jersey: John Wiley & Sons;

11. Chan, C.C., and R.S. Chen. (2005), "Using data mining technology to solve classification problems: a case study of campus digital library". The Electronic Library 24 (3), pp. 307-321.

12. Marban, O, l .(2009), "Toward data mining engineering: a software engineering approach. Information Systems", 34(1), pp. 87-107.

13. Sharma, N. (2005)," Discovering knowledge with text mining", M.S. Thesis, Texas A&M University.

14. Berson, A; Smith, S & Therling, k. (2001)," Building data mining applications for CRM", Mc Grawhill.

15. Mousazadegan HA, Zegardi H, 1387, A new model to solve the balance problem of the cost-based montage line, International journal of engineering sciences of Iran University of Science and Industry, No.19.

16. Jiawei H. M, K. (2001)," Data Mining: Concept and Techniques", Morgan Kaufmann,

17. Nemati. H. R Charmion, B , Kara. H. (2004) – Univercity of North Carolina at Greensboro, USA," Privacy Implications of Organizational Data Mining"

18. Berry MJA and GS Linoff. (2004), "Data Mining Techniques: For Marketing, Sales, and Customer Support". John Wiley.

19. Adel A, Ahmadi P, Sebt M. Desining Model for choosing human resources with data mining approach. Journal of Iranian Technology 2010; 2(4): 5. [In Persian].

20. Taqavi M, Nobari N, (1385). "The application of the evolutional algorithms in data mining", International conference of the research methods in the sciences, technologies and engineering, Tehran, [In Persian].

21. Clifton, C & Thuraisingham, B. (2001), "Emerging standards for data mining' Computer Standards and Interfaces", vol. 23, pp. 187-93.

22. Jeffrey W. Seifert. (2004), "Data Mining: An Overview". Available on line At: Www.Fas.Org/ Irp/Crs/Rl31798.Pdf.

23. Ha'eri Mehrizi AA, (1382), Data mining: concepts and methods and applications, MSc thesis, Allameh Tabatabaei University, Tehran, [In Persian].

24. Klosgen, W., and May, M. (2002), "Census Data Mining: An Application." Working Paper.

25. Banerjee, K. (1998), "Is data mining right for your library?", Computers in Libraries 18 (10), pp. 28-31.

26. Guenther, Kom. (2002)," Building Digital Libraries: applying data mining principles to library data collection". Computers in Libraries April pp. 60-63.

27. Seifert. Jeffrey W. (2004), "Data mining: An Overview. Congressional Research Service".

28. Pons-Porrata, A., R. Berlanga-Llavori, and J. Ruiz-Shulcloper. 2007. Topic discovery based on text mining techniques, Information Processing and Management 43 (3), pp. 752-768.

29. Chen H, Fuller SS, Friedman C, Hersh W. (2005), "Medical Informatics: Knowledge Management And Data Mining in Biomedicine". New York: Springer.

30. Englebardt SP, Nelson R. (2002), " Health-care informatics: an interdisciplinary approach". Philadelphia: Mosby, p. 125.

31. Canlas RD. (2009)," Data Mining in Healthcare: Current Applications and Issues [Online]. [cited 9 Aug 2009]; Available from: URL.

32. LaTour KM, Eichenwald S. Health Information Management: Concepts, Principles, and Practice. Chicago: AHIMA; 2002. p. 478-80.

33. Sepehri MM, Rahnama P, Shadpour P, Teimourpour B.(2009), "A data mining based model for selecting type of treatment for kidney stone patients". Tehran University Medical Journal; 67(6): 7-421[In Persian].

34. Balib RK. Clinical Knowledge Management: Opportunities and Challenges. Hershey: Idea Group Inc (IGI); 2005.

35. Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004; 25(8): pp. 5-69.