Data Mining, an Approach for Developing the Health Domain

Shahram Tofighi1, Ali Ghazvini2, Gholamhossein Pourtaghi3, Mansour Esmaeilpour4, Reza Shahhoseini5

Abstract

Nowadays, data mining as the process of arrangement and classifications of voluminous data is one of the most important techniques 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 techniques of data mining as one of the most important approaches in the development of health domains.

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 information 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 information banks by governmental offices and private 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 evolutorial trend of information technology which this evolutorial trend is arising from an evolutorial 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 techniques. 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 centered 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 Naibitt 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
Data mining has slowly but progressively been employed to the health domains, the progressive growth of this science in health fields are: the sensitivity of the science of medicine and its entwining with human 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 patterns can be resulted to change the equilibrium between death and life), amazement in the definition of data mining patterns can be resulted to change the equilibrium between death and life), amazement in the definition of data mining patterns can be resulted to change the equilibrium between death and life).
It can be said that changing the habit of the care providers can significantly improve their care quality to prevent future problems. This will help them recognize the unsafe patients and improve their rankings. The use of data mining techniques is among the most important applications of data mining in the health domain, one can refer to the following:

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 techniques 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 techniques is among the standard methods. 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.

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

(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.

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:

- 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 techniques 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.

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.

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 and helps diagnosis, treatments and drug manufacturing by analyzing the images, cardiograms and other clinical observations. 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.

Fuzzy logic: The fuzzy logic is more flexible in compared to the other techniques 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.
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
2. Daniel T. Larose, (2004). "Discovering Knowledge in Data: An Introduction to Data Mining".
15. Mousazadegan HA, Zeghari D, 1387, A new model to solve the balance problem of the cost-based montage line,
16. Jiawei H. M, K. (2001)," Data Mining: Concept and Techniques", Morgan Kaufmann,
20. Taqavi M, Nobari N, (1385), "The application of the evolulutional algorithms in data mining", International conference of the research methods in the sciences, technologies and engineering, Tehran, [In Persian].
23. Ha'er Mehrizi AA, (1382), Data mining: concepts and methods and applications, MSc thesis, Allameh Tabatabaei University, Tehran, [In Persian].
34. Balib RK. Clinical Knowledge Management: Opportunities and Challenges. Hershey: Idea Group Inc (IGI); 2005.