Data Mining: A Novel Outlook to Explore Knowledge in Health and Medical Sciences

Golbarg Mehrpoor¹, Mohammad Mehdi Azimzadeh²*, Amirhossein Monfared²

1. School of Medicine, Alborz University of Medical Sciences, Karaj, Iran
2. Department of Industrial Intelligence Research Group, ACECR, Zanjan Branch, Zanjan, Iran

*Corresponding Author: Mohammad Mehdi Azimzadeh, Software Engineering, School of Medicine, Alborz University of Medical Sciences, Karaj, Iran
Email: mehdiazimzadeh19@yahoo.com

Abstract

Today medical and Healthcare industry generate loads of diverse data about patients, disease diagnosis, prognosis, management, hospitals' resources, electronic patient health records, medical devices and etc. Using the most efficient processing and analyzing method for knowledge extraction is a key point to cost-saving in clinical decision making. Data mining, sometimes called data or knowledge discovery, is the process of analyzing data from different perspectives and summarizing it into useful information. In medicine, this process is distinct from that in other fields, because of heterogeneity and voluminosity of the data. Herein we reviewed some of published articles about application of data mining in several fields in medicine and healthcare.

Keywords: Medicine, Knowledge, Medical Informatics, Data Mining

Article History: Received: 18 Jun 2013 Revised: 1 Apr 2014 Accepted: 15 Apr 2014

Cite this article as: Mehrpoor G, Azimzadeh MM, Monfared A Data Mining: A novel outlook to explore knowledge in medical and health sciences Int J Travel Med Glob Health. 2014;2(2):87-90.

Introduction

Given the rapid and extensive development of science and technology, there is an urgent need to organize the information. In medical field, information system can considerably facilitate utilization of the information. Many sectors, nowadays, deal with electronic and non-electronic data, which can be converted into information and knowledge by employing different technologies. Information technology was introduced in 1960 to convert preliminary data of information systems and databases. In those early days, systems such as database management system were only in charge of rapidly saving great deal of data. However, technological development did not stop there and the data were gradually converted into useful knowledge [1]. Data in the age of information is considered as a valuable asset. Hence, development of an organization is highly dependent on available information and the way such information is analyzed. Evidently, many organizations have initiated quality improvement programs under different titles. Such programs can be more fruitful if they enjoy better analysis of the stored data. It is notable, however, that interpretation of enormous deal of data is not an easy job. Data mining is one of the methods that help us in this regard.

Data mining is about inferring or extracting knowledge out of an enormous deal of data available. The term comes from "Gold Mining" which refers to extracting gold from mountains [2, 3]. The science of data mining employs intelligent techniques to extract knowledge out a set of data. Such knowledge is not achievable using conventional statistic techniques. The science has the same meaning with "knowledge discovery in database" or KDD [4]. Data mining was introduced in the late 1990s and with its rapid growth we can expect big changes in knowledge generation in the world.

A workshop of extracting knowledge out of databases was held from 1989 to 1994. The term “data mining” was first used by Fiaz et al. Since then, the idea has been seriously studied in statistics [5], and first data mining journal was published in 1996 [6].

Five key features of data mining:

Extracting, changing, and opening transaction data in a data storage system.
Saving and managing data in multi-dimensional information bank system.
Providing data access for commercial analyzers and experts of information.
Data analysis using applications;
Introducing a useful frame of data such as graphs or tables [7]. There are three general types of data mining including clustering, classification, and association rule discovery [9].
A set of data, under cluster method, is clustered based on similarities and differences so that data in one cluster are similar and different from the data in other clusters. In general, clustering is the first step in data mining and put the data in the pertinent records for further analyses. The main clustering methods are classification, hierarchical, and density based [9].
Classification is a learning process under supervision, which takes place in two stages. First, a set of data are employed to create a model of data. The model actually describes the concepts and features of a set of data on which the model is built. The second stage is to implement or utilize the created model of the data on the data including all features thereof [10]. There are different algorithms and methods proposed for classification; for instance, decision tree, Bayes classifier, SVM, classification by neural networks, and rule based classification [10].

Rules of association try to find elements and items in a set of data that commonly occur in the data so that an associations between occurrences of data is assumed [10].

**Data mining applications**

1. Retailing: some of classic applications of data mining are:
   - Determining customers’ purchase pattern;
   - Analyzing market purchase portfolio;
   - Forecasting purchase through post (e-sale)
2. Insurance
   - Analyzing claims
   - Forecasting policy purchases
3. Banking:
   - Forecasting credit cards fraud
   - Determining fix number of customers
   - Determining popularity of credit cards based on social classes;
4. Space and space trips
   - Space information processing;
   - Space ship information processing;
   - Providing knowledge to make final decision to launch the space ship.

**Data mining and cancer**

Data mining nowadays is one of the best ways to diagnose and treat cancer as well as early diagnosis of it [11]. Artificial neural network (ANN) is one the most effective data mining methods and an approved technique to forecast survival of thyroid cancer patients [12]. Delen et al, used ANN, decision tree, and logistic regression to improve breast cancer forecast. They used the decision tree for extracting knowledge out of the data [13]. Landing et al, employed ANN and logistic regression to produce 5, 10, and 15 years forecasts of breast cancer patients. They used size of tumor, lymphatic node status, type of tissue, formation of tubolos, tumor necrosis and age as input variables and concluded that clustering trees and logistic regression are more effective for clinical interpretations [14]. Tolouie et al studied recurrence of breast cancer and supported vector machines (SVM) for forecasting recurrence of breast cancer, minimum error, and maximum accuracy [15].

In addition, ANN has been used for forecasting survival of esophageal cancer patients in [16], mortality rate among the liver hepatocellular carcinoma patient [17], and awareness among the liver cancer patients [18]. ANN outperformed statistical methods and TNM staging system in determining stage of different types of cancers [19].

**Data mining and diabetes**

Diabetes is a chronic and complicated disease with several symptoms. Industrialization of human life has brought increase in diabetes cases. About 200 million diabetics live in the world and 2 million of them are in Iran.

Miak et al, used card method to study the factors in development of diabetic symptoms [20]. Regression method was employed by Rolfling et al, to examine the relationship between blood sugar of diabetic I and HbA1c [21]. Big Hoan Cho et al, used SVM through feature section and visualization to find neuropathy among the diabetic [22].

ANN, decision tree, and logistic regression were compared regarding diagnosing diabetic or prediabetic patients among individuals with factor risk and the results showed that decision making tree had the highest susceptibility (75.13) and accuracy (77.87), and ANN had the lowest accuracy (73.23) [23].

Kim et al, used data mining (a priori algorithm) to survey the co-diseases and symptoms of diabetes and relationship between these, among 411414 patients [24]. Gregori et al, used data mining for monitoring diabetic patients [25]. Several studies have used different data mining algorithms to diagnose, manage and follow up diabetic patients. Ameri et al, used data mining algorithm (decision tree C5.0 and ANN) to classify diabetic patients based on their symptoms. They obtained the best results from the tree algorithm with accuracy of 89.06% and authenticity of model of 89.74% [26].

**Data mining and renal diseases**

Sepehri et al, employed decision tree to determine ureter stone treatment and argued that the model helped more patients to reach complete treatment [27]. Data mining techniques are mostly used for dialysis patients. Different available data mining techniques such as decision tree, fuzzy algorithm and so on are used in management and making decisions pertinent to dialysis patients and determining risk of cardiovascular diseases. In addition, these techniques are helpful in detecting early failure of venous fistul [28-30]. Decision tree has been showed to be more effective in follow up of the kidney implantation patients and the effects of risk
factors [31].

Data mining and cardiovascular diseases
Given the gravity of cardiovascular diseases - as the first cause of death in modern society - early diagnosis of the diseases is vital. Dehghani et al, used clustering data mining techniques to detect and forecast heart attacks [32]. Another study compared conventional data mining techniques such as decision-making tree, simple ANN Bayes, nearest neighbor K, and decision-making list and modern data mining tools such as weight association classifier (WAC) to achieve a proper algorithm to ensure accuracy of heart disease forecast. The results showed higher performance of simple ANN Bayes and decision tree [33]. Heart diseases analysis using evidence-based data mining techniques also known as Dampster-Shafer theory was also subject matter of another study. Austin et al, compared regression tree, ensemble-based methods, and conventional logistic regression to determine short-term (1 month) mortality rate among serious heart attack patients, and heart congestive failure. They concluded that boosted regression trees, trees and random forests outperformed conventional regression tree [34].

Data mining in health and hygiene
Data mining is also very useful in health and hygiene fields [35]. Electronic medical files can encompass plenty of data about symptoms, treatment, laboratory and medical results and so on. Valuable information can be extracted from these data [36-38]. In addition, the information can be helpful for hospital infection controls [39], ranking hospitals [39], and implementation of health services [40-41].

Conclusion
What mentioned above was few cases of data mining applications in providing health services to cancer, diabetic, kidney, and cardiovascular patients and health and hygiene services. Over the 20 years since the introduction of the concept, the applications have grown rapidly in medical and health fields. Apparently, data mining technique will extend to all fields of health and become one of the main decision making instruments in diagnosis, treatment, and health policy making. It is recommended therefore physicians, researchers, and decision makers put more emphasis on data mining concepts along with statistics and health system management techniques.

Acknowledgments
Author thank from staff of Industrial Intelligence Research Group.

Authors’ Contribution
All authors had equal duties in this article.

Financial Disclosure
The authors declared no financial disclosure.

Funding/Support
This work did not receive any funding or support.

References
11- Asadi N, Sadrodi M. Employing data mining to identify cancer risk factors and determine the optimal treatment in Namazi hospital cancer database. 16th Annual National Conference of Computer Society of Iran, 2010; Sharif University.


32. Dehghan T, Afshari Saleh M, Khalilzadeh M. A genetic K-means clustering algorithm for heart disease data. 5th Conference of Data Mining of Iran, 2011; Amirkabir University.

33. Zamanpoor S, Shamsi M. Assess and compare the accuracy of data mining algorithms to predict a heart disease. 4th Iranian Conference on Electrical and electronics Engineering, 1391 Gonabad, Iran.


