Binary Journal of Data Mining & Networking 4 (2014) 29-32

http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

Survey: Data Mining Techniques in Medical Data Field

Shiv Shakti Shrivastava[1] , Anjali Sant[2]

*Corresponding author:

Shiv Shakti Shrivastava

[1]Mewar University, Chittorgarh, Rajasthan. India
[2]Dept of Mathematics, BIST, Bhopal, India

Abstract

Now days most of the research area are working on data mining techniques in medical data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, decision tree, machine learning algorithms like Support Vector Machines, Principle Component Analysis etc., Rough Set Theory and Fuzzy logic. In particular we consider a case study using classification techniques on a medical data set of diabetic patients.

Keywords : Healthcare, health data, medical diagnosis, data mining, knowledge discovery in databases (KDD)

Introduction

Problem of showing uncertain data in economics, engineering, environmental science, sociology, medical science, and many other fields are very important for solving practical problems [1]. The evolution of the mass of data and number of existing databases far exceeds the skill of humans to analyze this data, which makes both a need and an opportunity to extract knowledge from databases. Medical databases have gathered large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge.

It is well known that in Information Technology (IT) driven society, knowledge is one of the most significant assets of any organization. The role of IT in health care is well established. Knowledge Management in Health care offers many challenges in creation, dissemination and preservation of health care knowledge using advanced technologies. Pragmatic use of Database systems, Data Warehousing and Knowledge Management technologies can contribute a lot to decision support systems in health care.

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows:

"Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data"[8].

Data mining technology provides a user - oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Following are some of the important areas of interests where data mining techniques can be of tremendous use in health care management.

Traditionally, decision making in health care is based on the ground information, lessons learnt in the past resources and funds constraints. However, data mining techniques and knowledge management technology can be applied to create knowledge rich health care environment.

Motivation

The main challenges in medical data classification are its huge size and its vagueness. Many researchers are trying to achieve the higher accuracy for medical data classification by reducing or selecting the informative data among the thousands of data. For cancer classification, researchers have used some machine learning algorithms like Support Vector Machines, Principle Component Analysis etc., Rough Set Theory and Fuzzy logic. In this paper we have applied bijective soft set based for classification for breast cancer, Pima Indians. However there can be a concern of patient privacy. It is more than clear that the role of data mining is not to practice medicine but to improve useful information and knowledge so that better treatment and health care be provided.

Data Mining & Techniques

Knowledge discovery in medical databases: Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry [2,10]. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven. Data mining is a young interdisciplinary field closely connected to data

warehousing, statistics, machine learning, neural networks and inductive logic programming.

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 healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain.

Massive healthcare data needs to be converted into information and knowledge, which can help control, cost and maintains high quality of patient care. Healthcare data includes Patient centric data and Aggregate data.

For health care organization to succeed they must have the ability to capture, store and analyze data. Online analytical processing (OLAP) provides one way for data to be analyzed in a multi-dimensional capacity. With the adoption of data warehousing and data analysis/OLAP tools, an organization can make strides in leveraging data for better decision making [3].

Many healthcare organizations struggle with the utilization of data collected through an organization online transaction processing (OLTP) system that is not integrated for decision making and pattern analysis. For successful healthcare organization it is important to empower the management and staff with data warehousing based on critical thinking and knowledge management tools for strategic decision making. Data warehousing can be supported by decision support tools such as data mart, OLAP and data mining tools. A data mart is a subset of data warehouse.

It focuses on selected subjects. Online analytical processing (OLAP) solution provides a multi-dimensional view of the data found in relational databases. With stored data in two-dimensional format OLAP makes it possible to analyze potentially large amount of data with very fast response times and provides the ability for users to go through the data and drill down or roll up through various dimensions as defined by the data structure.

With the widespread use of medical information systems including databases, there is an explosive growth in their sizes, Physicians are faced with a problem of making use of stored data. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable. A Data Warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions[3]. A data warehouse is also often viewed as architecture constructed by integrating data from multiple heterogeneous sources to support structured and/or ad- hoc queries, analytical reporting and decision making.

Data mining techniques in health care: There are various data mining techniques available with their suitability dependent on the domain application. 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.

Classification data mining techniques: We now describe a few Classification data mining techniques with illustrations of their applications to healthcare.

Rule induction: is the process of extracting useful 'if-then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form.

K- Mean Algorithm

The k-means algorithm is a simple iterative method to partition a given dataset into a user- specified number of clusters, k. This algorithm has been discovered by several researchers across different disciplinesGray and Neuhoff [6] provide a nice historical Back ground for k-means placed in the larger context of hill-climbing algorithms. The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1, . . . , N }, where xi ∈ d denotes the ith data point. The algorithm is initialized by picking k points in d as the initial k cluster representatives or —centroids||. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times.

Association Rule Mining (Arm)

ARM is a popular data miring technique, which aims at discovering strong interesting patterns (associations) between items in vast data sets. For instance, X =$ Y, read "X implies YI',is an association rule, which is interpreted as "if X occurs it is most likely that Y also will occur". The task of mining asslociation rules is simply searching for items which occur frequently (large items), defined by a user-defined minimum frequency (minimum support), and then finding patterns in their occurrences. Various algorithms exist to efficiently search for and count large item sets. Hipp et al. [14] provide a detailed slulrey of ARM algorithms. A number of them have bottlenecks, which limit the size and nature of datasets they can efficiently mine. For instance, generating candidate items sets have always been a source of bottleneck in the implementation of Apriori-like ARM algorithms in large data sets. Another problem encountered is the I/O overhead, which occurs when the transaction database exceeds the size of main memory.

Rule Induction Method has the potential to use retrieved cases for predictions. The following example gives rule induction method for prediction blood alcohol concentration. This is only for example purpose.

Table 1: Attributes for alcohol measurement

Attributes

Age (in yrs)	Meal (empty stomach, lunch, full)
Sex (M/F)	Amount of alcohol (ethanol in units)
Mass (in kg)	Blood_alcohol content (high/low)
Tobacco_use	Blood_pressure(high/low)
Height (in cm)	Time_duration (time spent drinking)

Decision tree: It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree-based models which include classification and regression trees, are the common implementation of induction modeling[5]. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT[5].

The decision tree is built from the very small training set. In this table each row corresponds to a patient record. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

References

[1]. Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[2]. Glymour C, Madigan D, Pregidon D, and Smyth P. 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.

[3]. Shams K. and Frashita M. 2001. Data Warehousing Toward Knowledge Management. Topics in Health Information Management, 21: 3.

[4]. Jones AW. 1990. Physiological Aspects of Breath-Alcohol Measurements. Alcohol Drugs Driving, 6:1-25.

[5]. Han J, and Kamber M. 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

[6]. Lu H,. Setiono R, and Liu H. 1996. Effective data mining using neural networks. IEEE Trans. On Knowledge and Data Engineering, 5: 8.

[7]. Miller A, Blott B, and Hames T. 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464.

[8]. Miller A. 1993. The application of neural networks to imaging and signal processing in astronomy and medicine. Ph.D. Thesis, Faculty of Science, Department of Physics, University of Southampton.

[9]. Weinstein J, Kohn K, and Grever M. *et al.*, 1992. Neural computing in cancer drug development: Predicting mechanism of action. Science, 258: 447-451.

[10]. Stanfford GC, Kelley PE, JSyka EP, Reynolds WE, and Todd JF. 1984. Recent improvements in and analytical applications of advanced ion-trap technology. Intl. J. Mass Spectrometry Ion Processes, 60: 85-98.

[11]. Robinson PJ. 1997. Radiology's Achilles's heel: Error and variation in the interpretation of the Roentgen image. Radiol. Brit. J.

Itchhaporia D, Snow PB, Almassy RJ, and Oetgen WJ. 1996. Artificial neural networks: Current status in cardiovascular medicine.

[12]. Schnorrenberg F, Pattichis CS, Schizas CN, Kyriacou K, and Vassiliou M. 1996. Computer aidded classification of breast cancer nuclei.

[13]. Choi HK, Jarkrans T, Bengtsson E, Vasko J, Wester K, Malmsrom PU, and Bausch C. 1997. Image analysis based carcinoma. Comparison of object, texture and graph based methods and their reproducibility.

[14]. Simon BP, and Eswaran C. 1997. An ECG classifier designed using modified decision based neural networks.

[15]. Romeo M, Burden F, Quinn M. Wood B, and McNaughton D. 1998. Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer.

[16]. Ball G, Mian S, Holding F. Allibone Ro *et al.*, 2002. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomakers. Bioinformatics, 18: 395-404.

[17]. Aleynikov S, and Micheli-Tzanakou E. 1998. Classification of retinal damage by a neural network based system.

[18]. Domine D, Guillon C, Devillers J, Lacroix J, and Dore JC. 1998. Non linear neural mapping analysis of the adverse effects of drugs.

[19]. Sharma A. and Roy RJ. 1997. Design of a recognition system to predict movement during anesthesia. IEEE Transactions

[20]. Aboul Ella HAssanien. "Intelligent data analysis of breast cancer based on Rough set theory," International Journal on Artificial Intelligence Tools ,vol. 12, no. 4, 2003, pp. 465-479.

[21]. Akta H, Çagman N. "Soft sets and soft groups," Information Science, 177 (13), 2007, pp. 2726–2735.

[22]. Gong K. et al.,"The bijective soft set with its operations," Computers & Mathematics with Applications 60, 2010.

[23]. HerawanT, Deris MM. "A direct proof of every rough set is a soft set," Third Asia International Conference on

Modelling & Simulation, pp. 119–124, 2009.

[24]. Ivo D, and Gunther G. "Statistical Evaluation of Rough Set Dependency Analysis," International Journal of Human-Computer.Studies, vol. 46, pp. 589- 604, 1997.

[25]. Kalaiselvi N, Hannah Inbarani H. "Fuzzy Soft Set Based Classification for Gene Expression Data,".

[26]. Maji PK, Biswas R and Roy AR. "Fuzzy Soft Sets," Journal of Fuzzy Mathematics, vol. 9, pp. 589-602, 2001.

[27]. Maji PK, Roy A, Biswas R. "An application of soft sets in a decision making problem," Computer Mathematical Application, vol. 44, pp.77– 1083, 2002. Studies, vol. 46, pp. 589- 604, 1997.

[28]. Kalaiselvi N, Hannah Inbarani H. "Fuzzy Soft Set Based Classification for Gene Expression Data,".

[29]. Maji PK, Biswas R and Roy AR. "Fuzzy Soft Sets," Journal of Fuzzy Mathematics, vol. 9, pp. 589-602, 2001.

[30]. Maji PK, Roy A, Biswas R. "An application of soft sets in a decision making problem," Computer Mathematical Application, vol. 44, pp.77– 1083, 2002.

[31]. Molodtsov D. "Soft set theory-first results," Computers & Mathematics with Applications, vol. 37, pp. 19–31, 1999.

[32]. Pei D Miao D. "Soft sets to information systems," Granular Computing, IEEE International Conference, 2005.