Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 4, Issue 4; July-September, 2017, pp. 253-257

http://www.krishisanskriti.org

Studyon Data Mining Techniques and Applications in Healthcare

Sharad Mathur[1] and Bhavesh Joshi[2]

PAHER University, Udaipur (Raj.)
E-mail: [1]sharad_mathur2002@yahoo.com, [2]bablajoshi@gmail.com

Abstract—*Data Mining is one of the most promising areas of research and now a days it becomes increasingly popular in health organization. Data Mining plays an important role for uncovering new trends in healthcare sector which in turn helpful for all the parties associated with this field like doctors, patients, pharmacy industries, health insurance companies etc.*

This paper explores the various Data Mining techniques such as classification, clustering, association and regression. In this paper, we present a brief introduction of these techniques and their advantages and disadvantages. This paper also highlights applications of Data Mining in healthcare.

INTRODUCTION

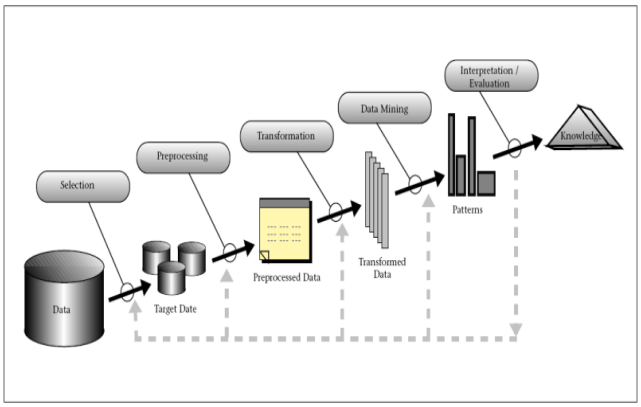
Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. In health industry, Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods [1].

The analysis of health data improves the healthcare by enhancing the performance of patient management tasks. The outcome of Data Mining technologies are to provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organization provides them effective treatments [2]. It can also useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management.

Hospitals must also minimize the cost of clinical tests. Theycan achieve these results by employing appropriatecomputer-based information and/or decision supportsystems. Health care data is massive. It includes patientcentric data, resource management data and transformeddata [3]. Health care organizations must have ability to analyzedata. Treatment records of millions of patients can be storedand computerized and data mining techniques may help inanswering several important and critical questions related tohealth care.

Knowledge discovery and data mining

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery[4]. Knowledge discovery as a process is depicted in Following Figure and consists of an iterative sequence of the following steps:



- Data cleaning (to remove noise and inconsistent data).
- Data integration (where multiple data sources may be combined).
- Data selection (where data relevant to the analysis task are retrieved from the database).
- Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
- Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
- Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures).

- Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

Data Mining techniques

TechniquesThere are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections.

Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That is the reason why association technique is also known as relation technique[5]. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics [6]. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay" [7]. And then we can ask our data mining software to classify the employees into separate groups.

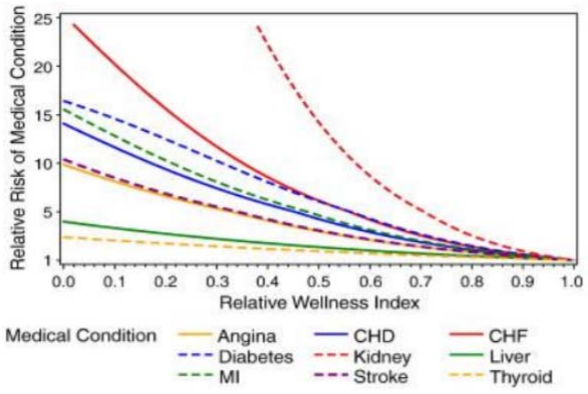
Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes [8]. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library [9].

Regression

Regression is used to find out functions that explain the correlation among different variables. A mathematical model is constructed using training dataset. In statistical modeling two kinds of variables are used where one is called dependent variable and another one is called independent variable and usually represented using 'Y' and 'X'. There is always one dependent variable while independent variable may be one or more than one. Regression is a statistical method which investigates relationships between variables. By using Regression dependences of one variable upon others may be established [21].

Regression is widely used in medical field for predicting the diseases or survivability of a patient. Figure below represents an application of logistic regression for the estimation of relative risk for various medical conditions such as Diabetes, Angina, stroke etc [22].



Example of Logistic Regression

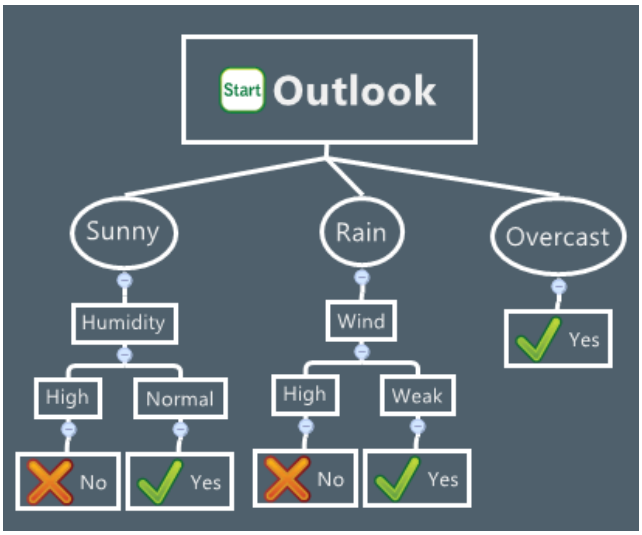
Prediction

The prediction, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Decision trees

The A decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers [10]. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. For example, We use the

following decision tree to determine whether or not to play tennis:



Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the week. And if it is sunny then we should play tennis in case the humidity is normal.

Support Vector Machine (SVM)

Support vector machines (SVM) are a classification technique originated from statistical learning theory.Depending on the chosen kernel, SVM selects a set of data examples (support vectors) that define the decision boundary between classes [11]. SVM have been proven for excellent classification performance, while it is arguable whether support vectors can be effectively used in communication of medical knowledge to the domain experts.SVMs are well suited to dealing with interactions amongfeatures and redundant features [12].

Applications of Data Mining in Healthcare

Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories.

Treatment effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.The use of classification algorithms to help in the early detection of heart disease, a major public health concern all over the world. The use of data mining as a tool to aid in monitoring trends in the clinical trials of cancer vaccines. By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data [13].

Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists [14].

Fraud and abuse

To identify fraud and abuse data mining applications often set up norms and then recognize unusual patterns of claims by physicians, clinics, laboratory or some others. These data mining applications can also throw a light on unsuitable prescriptions or referrals and false insurance and health claims [15].

Ranking Hospitals

Organizations rank hospitals and healthcare plans based on information reported by healthcare providers. There is an assumption of uniform reporting, but research shows room for improvement in uniformity. Data mining techniques have been implemented to examine reporting practices. With the use of International Classification of Diseases, 9th revision, codes (risk factors) and by reconstructing patient profiles, cluster and association analyses can show how risk factors are reported [16]. Standardized reporting is important because hospitals that underreport risk factors will have lower predications for patient mortality. Even if their success rates are equal to those of other hospitals, their ranking will be lower because they reported a greater difference between predicted and actual mortality[16]. Standardized reporting would also be important for meaningful comparisons across hospitals.

Cost effective treatment

Data mining allows organizations and institutions to get more out of existing data at minimal extra cost. KDD and data mining have been applied to discover fraud in credit cards and insurance claims [15]. By extension, these techniques could also be used to detect anomalous patterns in health insurance claims, particularly those operated by PhilHealth, the national healthcare insurance system for the Philippines.

Evidence-based medicine and prevention of hospital errors

When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. For instance, an ongoing study on hospitals and safety found that about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including

doctors) been more careful in avoiding errors [17].By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

Medical Device Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless biosensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients[18]. Ubiquitous Data Stream Mining (UDM) techniques such as light weight, one-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge and available memory.

Pharmaceutical Industry

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making.

Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized[19]. Three layers of hospital management:

- Services for hospital management
- Services for medical staff
- Services for patients

System Biology

Biological databases contain a wide variety of data types, often with rich relational structure. Consequently multirelational data mining techniques are frequently applied to biological data [20]. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

Conclusion

Over the years data mining has enjoyed tremendous success, the application areas expanded continuously and the mining techniques also kept up improving. Various problems have appeared and have been solved by data mining researchers. Number of techniques of data mining is being used by various researchersin healthcare sector and this paper has given focus on these techniques. This paper also summarized application of data mining for healthcare sector.

References

[1] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).

[2] C. T. Su, P. C. Wang, Y. C. Chen, and L. F. Chen, "Data mining techniques for assisting the diagnosis.

[3] Cheon-Pyo Lee and Jung P Shim "An exploratory study of radio frequency identification (RFID) adoption in the healthcare industry", European Journal of Information Systems, Vol. 16, pp. 712–724, 2007.

[4] Fayyad, U, Piatetsky-Shapiro, G and Smyth, P, 1996, ``From data mining to knowledge discovery" AI Magazine 17(3) 37-54 Cristianini N, Shawe-Taylor J. 2006. an introduction to support vector machines and other kernel- based learning methods. Cambridge Univ. Press

[5] T. P. Hong, K. Y. Lin, and S. L. Wang, "Fuzzy data mining for interesting generalized association rules," Fuzzy Sets Syst., vol. 138, no. 2, pp. 255–269, 2003.

[6] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST ISSN: 2229- 4333, vol. 2, no. 2, (2011) June

[7] P. R. Harper, "A review and comparison of classification algorithms for medical decision making", Health Policy, vol. 71, (2005), pp. 315-331.

[8] P. Berkhin, "A Survey of Clustering Data Mining," Group. Multidimens. Data, no. c, pp. 25–71, 2006.

[9] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3918 LNAI, pp. 199–204, 2006.

[10] Apte& S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsaptewe issue_with_cover.pdf, (1997).

[11] Cristianini N, Shawe-Taylor J. 2006. an introduction to support vector machines and other kernel- based learning methods. Cambridge Univ. Press.

[12] Vapnik VN. 1998. Statistical Learning Theory,wiley.

[13] Cao, X., Maloney, K.B. and Brusic, V.(2008). Data mining of cancer vaccine trials: a bird's-eye view. Immunome Research, 4:7. DOI:10.1186/1745-7580-4-7.

[14] HianChye K, Gerald T. 2005, Data mining applications in healthcare, Journal of healthcare information management: JHIM.19 (2): 64-72.

[15] Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). Survey of fraud detection techniques. In Networking, Sensing and Control, 2004 IEEE International Conference on Networking, Sensing and Control. (2) 749-754.

[16] Cerrito P. Using text analysis to examine ICD-9 codes to determine uniformity in the reporting of MedPAR data. Presented at the Annual Symposium of the American Medical Informatics Association; November 9-13, 2002; San Antonio, TX.

[17] Health Grades, Inc. (2007). The Fourth Annual HealthGrades Patient Safety in American Hospitals Study.

[18] Mobile Data Mining for Intelligent Healthcare Support.

[19] ShusakuTsumoto and Shoji Hirano, ―Temporal Data Mining in Hospital Information Systems.

[20] J. Bockhorst, M. Craven, D. Page, J. Shavlik, and J. Glasner. A Bayesian network approach to operon prediction. Bioinformatics, 19(10):1227--1235, 2003.]

[21] J. Fox, "Applied Regression Analysis, Linear Models, and Related Methods", (1997).

[22] C. Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.