International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Data Mining Techniques in Healthcare Industry

Mahak*

Department of CSE, Kurukshetra University Kurukshetra, India

Accepted 12 Feb 2017, Available online 23 Feb 2017, **Vol.7, No.1 (Feb 2017)**

Abstract

Data Mining has an essential & vital role now days. It has been used intensively and broadly by several organizations such as e-business, marketing and retail because of which it is now applicable in knowledge discovery in databases (KDD) in many industrial areas and economy. Data mining is greatly gaining its importance and usage in the healthcare industry. Patient centric data, their treatment data and resource management data all are included in Health Care industry. It has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data. Advanced data mining techniques can be used to discover hidden patterns on healthcare data. This paper includes the investigation of present methods of KDD and briefly examines the prospective use of classification based data mining techniques such as decision tree and Artificial Neural Network to enormous volume of healthcare data.

Keywords: *Data Mining, Healthcare, Knowledge Discovery in Databases (KDD), Decision tree, Artificial Neural Network.*

1. Introduction

In today's information time, there is a need for a powerful analytical solution for the extraction of the useful information from the large amount of data collected and stored in an organization's databases or repositories. 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 (Harleen Kaur et al, 2006).

In healthcare, data mining is becoming gradually more well-liked, if not ever more essential. Several factors have motivated the use of data mining applications in healthcare (Shelly Gupta *et al* ,August 2011)(Witten *et al*). The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders (Hand DJ *et al*).

Data mining, or knowledge discovery, is the computer assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data (Dr .M Hemalatha *et al*, 2011) Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions (Shams *et al*) Data mining expertise provide a consumer leaning approach to new and unknown patterns in the data. The exposed knowledge can be used by the healthcare administrators to progress the superiority of service.

Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection (Witten *et al*) (Li wanqing *et al*) Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data (Jans *et al*) Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data (Silver *et al*) (Lin J *et al*, 2003). Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care (Major *et al*,2002).

This paper is organized as follows: Section 1 introduction and describe importance of the data mining in the healthcare industry. Section 2 & 3 describes the data mining concepts and Knowledge Discovery. In section 4 describes the data mining techniques in healthcare. Section 5 explains the importance and uses of data mining in medicine. Section 6 concludes the paper.

2. Data mining concept

Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends

*Corresponding author is a PhD Scholar

that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data.

These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as:

- Forecasting: Estimating sales, predicting server loads or server downtime.
- Risk and Probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes.
- Recommendations: Determining which products are likely to be sold together generating recommendations.
- Finding Sequences: Analyzing customer selections in a shopping cart, predicting next likely events.
- Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities.

3. Data mining and knowledge discovery

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The figure 1.1 shows data mining as a step in an iterative knowledge discovery process (Fayyad *et al*,1996).

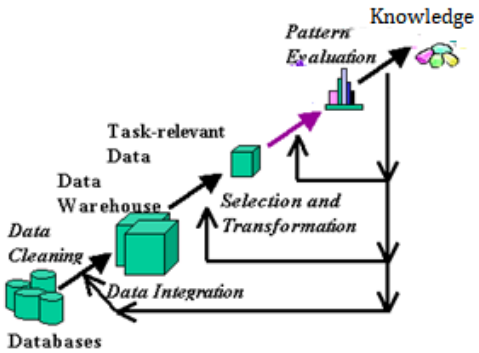

Fig.1 Date Mining is the core of knowledge process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data Cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data Integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data Selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data Transformation**: also known as data consolidation, it is a phase in which selected data is transformed into forms appropriate for the mining procedure.
- **Data Mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern Evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge Representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called gold mining and not rock mining, thus by analogy, data mining should have been called knowledge mining instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

4. 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

IF conditions THEN conclusion

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 (Hans *et al* ,2001) Decision tree models are best suited for data mining.

Artificial Neural Network (ANN): is a collection of neuron –like processing units with weight connections between the units. These models mimic the human brain and learn the patterns of a data set in order to make predictions. Artificial Neural Networks (ANN) are analytical techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from previous observations after executing a process called learning from existing data
(Lu H *et al*,1996)

Artificial Neural Network is one of many data mining analytical tools that can be utilized to make predictions on key healthcare indicator such as cost or facility utilization. Neural networks are known to produce highly accurate results and in medical applications, can lead to appropriate decisions.

5. The importance and uses of data mining in medicine and public health

The demand and want for data mining is more in field of healthcare, regardless of variations and conflicts in processes. Various discussions led to the demand of data mining in the field of healthcare which includes both public health as well private health. Many facts can be achieved from the past data stored in computers. Since the data is huge in quantity so it's a disadvantage for persons to examine the entire data and gain awareness (Chang *et al* ,2006). Specialists consider that the improvement in medicals has reduced leading to complication of recent medical data. To overcome this drawback computers and data mining can be utilized.

Evidence-based medicine and prevention of hospital errors: On utilizing data mining on the available data much new informative and possibly life-rescuing information is achieved or else which would have left unutilized. For example, in recent research on hospitals and wellbeing it was originated that almost 87% of the deaths in the United States could have been reduced if the errors would have been lowered by the hospital staff (Health Grade ,2007).The preventive measurements could have been adopted by the hospitals and government supervisors using data mining to hospital data.

Policy-making in public health: The utilization of data mining in healthcare data helped health centers to determine methods that would lead to policy suggestions to the Public Health Institute. Decision making can be improved by proper utilization of data mining & decision support techniques (Pogorelc *et al*,2010).

More value for money and cost savings: Extra information can be obtained at less additional price by the firms and institutions using data mining. To know the information about scam in credit cards and insurance claims KDD & data mining is utilized (Nada Lavvac *et al*).

Early detection and/or prevention of diseases: To obtain the before time recognition of heart problem which is the main public health issue through the world, Cheng et al mentioned the application of arrangement strategy.

Early detection and management of pandemic diseases and public health policy formulation: For before time identification and supervision of pandemics health specialists have preferred to utilize data mining. To obtain the reasons behind occurrence of diseases (Bailey –Kellog *et al*) discussed various methods which is a mixture of spatial modeling, simulation and spatial data mining. The output of examined data mining in the simulated situation can be utilized further to discover and organize disease causes.

Conclusion

In this paper describe importance and uses of data mining concept in the healthcare system. The focus of the study was to discuss the various data mining techniques such as rule induction techniques, decision tree and Artificial Neural Networks which can support healthcare system via generating strategic information. There are different data mining techniques that can be used for the identification and prevention of cardiovascular disease among patients. In future we intend to improve performance of these basic classification techniques by creating meta model which will be used to predict cardiovascular disease in patients.

References

Harleen Kaur; Siri Krishan Wasan(2006) Empirical Study on Applications of Data Mining Techniques in Healthcare Journal of Computer Science 2 (2): pp 194-200

Shelly Gupta; Dharminder Kumar (August 2011) ; Anand Sharma(August 2011) Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data, International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4.

Witten; I. H. Frank, E. (2000). Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA. USA, 371 pp.

Hand D.J((2001); Mannila H. Smyth P. (2001). Principles of data mining, MIT Press, Boston, MA.,USA.

M.Hemalatha; S.Megala (31th march 2011) Mining Techniques in Health Care: A Survey of Immunization journal of theatrical and applied Information Vol.25 no.2.

Shams K (2001));M. Frashita, (2001). Data Warehousing Toward Knowledge Management. Topics in Health Information Management, Vol 21: 3.

Banu Rahaman; Shashi M(2010), Sequential mining equips e-Health with knowledge for managing diabetes, 4th International Conference on New Trends in Information Science and Service Science (NISS), pp.65-71.

Li wanqing; Ma lihua, Wei dong Wseas.(2010), Data Mining Based on Rough Sets in Risk Decision-making: Foundation and Application, transactions on computers Issue 2, Volume 9, pp: 113-121.

Jans, Nadine Lybaert; Koen Vanhoof.(2009), A Framework for Internal Fraud Risk Reduction at IT Integrating Business processes The International Journal of DigitalAccounting Research. Vol.9, pp.1-29.

Silver (2001); M. Sakata(2001); T. Su, H.C. Herman(2001); C. Dolins (2001); S.B. ; O'Shea, M.J. (2001). Case study: how to apply data mining techniques in a healthcare data warehouse, Journal of Healthcare Information Management, Vol.15(2)

Lin, J; Hwang, M. Becker, J. (2003): A fuzzy neural network for assessing the risk of fraudulent financial reporting, Managerial Auditing Journal, Vol. 188,pp 657-665.

Major, J; Riedinger, D. (2002), EFD: Ahybrid knowledge/statistical-based system for the detection of fraud, The Journal of Risk and Insurance, Vol. 693,pp 309-324

Fayyad; Piatetsky-Shapiro, Smyth(1996), Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, pp.1-34.

Han; J. M. Kamber, 2001. Data Mining Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

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

Cheng, T.H; Wei, C.P; Tseng, V.S. (2006) Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).

Health Grades, Inc. 2007. The Fourth Annual Health Grades Patient Safety in American Hospitals Study.

Pogorelc, B; Gams, M(Dec 2010); Medically Driven Data Mining Application: Recognition of Health Problems from Gait Patterns of Elderly, Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol, no., pp.976-980.

Nada Lavrac, Marko Bohanec, Aleksander Pur, Bojan Cestnik, Marko Debeljak, Andrej Kobler (2007): Data mining and visualization for decision support and modeling of public health-care resources. Journal of Biomedical Informatics 40(4): 438-447.

Bailey-Kellog, C. Ramakrishnan, N. and Marathe, M. Spatial Data Mining to Support Pandemic Preparedness. SIGKDD Explorations (8) 1, 80-82.