Applications of Data Mining Techniques: Empowering Quality Healthcare Services

Dr. Meenu Dave[1], Priyanka Dadhich[2]

ABSTRACT

Data pertaining to any industry is multiplying at an enormous speed and in healthcare industry the produced data is not only massive but also quite complex. As a result, proper handling of data is of prime importance in order to convert the available data into useful information that leads to knowledge and apposite decision making. Use of data mining in the healthcare industry is proving to be a boon for attaining speedy, accurate and futuristic results. This paper presents a systematic approach of data mining techniques in healthcare industry and a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use today in medical research and healthcare sector. This paper also focuses on some critical issues and challenges associated with the application of data mining in the profession of healthcare practices.

KEYWORDS

C4.5, Data Mining, Healthcare Sector, ID3, k-NN, SVM.

INTRODUCTION

Healthcare is an information-intensive sector. The need to develop and organize new ways of providing healthcare information, data and knowledge has been attended by major advances in information and communication technologies. These new technologies are speeding an exchange and use of data, information and knowledge and are eliminating geographical and time barriers. These processes highly accelerated medical informatics development. Nowadays medical informatics shows its significance as a multidisciplinary science developed on the basis of interaction of information sciences with medicine and health care in accordance with the attained level of information technology [1].

Today's healthcare environments use electronic health records that are shared between computer systems and which may be distributed over many locations and between organizations, in order to provide information to internal users, to payers and to respond to external requests. With increasing mobility of population, patient data is accumulating in different places, but it needs to be accessible in an organized manner on a national and even global scale.

There are following important areas of interests where data mining techniques can be of tremendous use in health care management.

Forecasting treatment costs and demand of resources
Identification of best practices in the treatments of specific diseases.

- Public Health Informatics
- e-governance structures in health care
- Data modeling for health care applications
- Executive Information System for health care
- Health insurance

In the current scenario, decision making system is based on data mining techniques and knowledge management technology that can be applied to create knowledge rich health care environment which was earlier based on the ground information, lessons learnt from the past resources and funds constraints [2].

FACTORS RESPONSIBLE FOR INCEPTION OF DATA MINING USE IN HEALTHCARE SECTOR

There are several factors which are responsible for the use of data mining in healthcare sector:

- First factor is the realization that data mining can generate information that is very useful to all parties involved in the healthcare industry. For example, data mining applications can help healthcare insurers detect fraud and abuse, and healthcare providers can gain assistance in making decisions, for example, in customer relationship management. Data mining applications can also benefit healthcare providers, such as hospitals, clinics and physicians, and patients, for example, by identifying effective treatments and best practices.

- 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. Fraud detection using data mining applications is prevalent in the commercial world, like, in the detection of fraudulent credit card transactions. Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection.

- Another factor is that the huge amounts of data generated by healthcare transactions are too complex and huge to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering

[1]*Assistant Professor, Department of Computer Science & IT, Jagan Nath University, Jaipur, India*
Email: [1] meenu.dave@jagannathuniversity.org
[2]*Research Scholar, Department of Computer Science & IT, Jagan Nath University, Jaipur, India*
Email: [2] abhipriya_dadhich@rediffmail.com

Copyright © IJICCT–JUL 2013;Vol 1,Issue 1;ISSN 2347-7202

This journal is cited as : **JIMS 8i-Int'l J. of Inf. Comm. & Computing Technology(IJICCT)**

patterns and trends in large amounts of complex data. 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. Insights gained from data mining can influence cost, revenue, and operating efficiency while maintaining a high level of care.

Apart from the above mentioned, a few more sub-factors are also worth mentioning:

A. Unsupervised Knowledge Discovery in Medical Databases

With increasing data in medical databases, medical data mining becomes an essential step in knowledge discovery. Some of these analyses use techniques from the machine learning literature, including propositional rules from databases using rough sets, implementation of these rules in an expert system, use of Bayes models to find similar cases, applying a finite-mixture-augmented naïve-Bayes model to classify cases and constructing decision trees and neural networks to classify cases [3,7].

B. Treatment Effectiveness

Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective.

It also has developed clinical profiles to give physicians information about their practice patterns and to compare these with those of other physicians and peer-reviewed industry standards **[5]**.

Other data mining applications related to treatments include associating the various side-effects of treatment, collating common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub-populations that respond differently from the mainstream population to certain drugs, and determining proactive steps that can reduce the risk of affliction.

C. Heart Diseases

Coronary Heart Disease (CHD) is a major cause of disability in adults in and common cause of death in Europe, USA, South Asia, etc., It has been predicted that all the regions of the world will be affected due to CHD by the year 2020. Coronary Heart Disease refers to the failure of coronary circulation to supply adequate circulation to cardiac muscle and its surrounding tissue [8].

This restricts the supply of blood and oxygen to the heart, particularly during exertion when the myocardial metabolic demands are increased. As the degree of coronary artery disease progresses, there may be near complete obstruction of the lumen of the coronary artery, severely restricting the flow of oxygen-carrying blood to the myocardium. Individuals with this degree of coronary artery disease typically have suffered from one or more myocardial infarctions (heart attacks), and may have signs and symptoms of chronic coronary ischemia, including symptoms of angina at rest and flash pulmonary edema [9].

Some of the data mining studies are already done by using the Random forest classification algorithm for evaluating and predicting various events related to CHD. The 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. Data warehousing can be supported by decision support tools such as data mart, online analytical processing (OLAP) which provides a multi-dimensional view of the data found in relational databases [3,10].

D. Hospital Infection Control

Nosocomial infections affect two million patients each year in the United States, and the number of drug-resistant infections has reached unprecedented levels. Early recognition of outbreaks and emerging resistance requires proactive surveillance.

Computer-assisted surveillance research has focused on identifying high-risk patients, expert systems, and possible cases and detecting deviations in the occurrence of predefined events.

A surveillance system that uses data mining techniques to identify new and interesting patterns in infection control data has been implemented at the University of Alabama. The system uses association rules on culture and patient care data obtained from the laboratory information management systems and generates monthly patterns that are reviewed by an expert in infection control.

Developers of the system conclude enhancing infection control with the data mining system is more sensitive than traditional infection control surveillance, and significantly more specific [11,12].

E. 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. Standardized reporting is important because hospitals that underreport risk factors will have lower

Copyright © IJICCT–JUL 2013;Vol 1,Issue 1;ISSN 2347-7202

This journal is cited as : **JIMS 8i-Int'l J. of Inf. Comm. & Computing Technology(IJICCT)**

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. Standardized reporting would also be important for meaningful comparisons across hospitals [13].

F. Identifying high Risk patients

American Healthways provides diabetes disease management services to hospitals and health plans designed to enhance the quality and lower the cost of treatment of individuals with diabetes. To augment the company's ability to prospectively identify high-risk patients, American Healthways uses predictive modeling technology. Extensive patient information is combined and explored to predict the likelihood of short- term health problems and intervene proactively for better short- term and long-term results.

A robust data mining and model-building solution identifies patients who are trending toward a high-risk condition. This information gives nurse care coordinators a head start in identifying high-risk patients so that steps can be taken to improve the patients' quality of healthcare and to prevent health problems in the future [14].

DIFFERENT DECISION TREE ALGORITHMS OF DATA MINING USED IN HEALTHCARE

A. ID3

The aim of ID3 algorithm is to construct a decision tree that on the basis of answers to questions about the non-category attributes predicts correctly the value of the category attribute. ID3 algorithm uses a fixed set of examples to build a decision tree and then uses this tree to classify given data samples.

Applications of ID3:

- ID3 algorithm is used in knowledge acquisition for tolerance design.
- This algorithm is applied to calculate logistic performance.
- This is applicable in the field of computer crime forensics.
- For Cancer detection Id3 is very useful.
- This is also helpful to diagnose the heart diseases.

Advantages of ID3:

- The Id3 is based on decision tree. So it is simpler to execute.
- A statistical property called Information Gain is used; gain measures how well a given attribute separates a training data into targeted examples.
- The one with the highest information is selected.
- It creates understandable prediction rules from the training data set.
- It also builds fastest and short decision tree.

B. C 4.5

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of Information Entropy. The training data is a set S = s1, s2... of already classified samples. Each sample si = x1, x2... is a vector where x1, x2...represent attributes or features of the sample. The training data is augmented with a vector C = c1, c2... where c1, c2... represent the class that each sample belongs to.

Applications of C4.5:

- The C4.5 is applicable in web-based learning system.
- This algorithm is used to handle continuous attributes eg. temperature.
- It also helps to handle training data with some missing values.
- C4.5 improves computational efficiency.

Advantages of C4.5:

- Handling both continuous and discrete attributes.
- Handling training data with missing attribute values.
- Handling attributes with differing costs.
- Pruning trees after creation.

C. k-NN

The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.

Applications of k-NN:

- Knn is used to diagnose the cancer using Fourier-transform infrared spectroscopy.
- This is applied in unbalanced data distribution.
- This algorithm is also applicable in Android applications.
- Knn is also useful in vowel recognition.

Advantages of k-NN:

- It is simple to implement.
- It works fast for small training set.
- Its performance asymptotically approaches the performance of Bayes Classifier.
- It does not need any retraining if the new training pattern added to the existing training set.
- The output of k-NN algorithm can be interpreted as a posterior probability of the input pattern belonging to a class thus the output provides the relative class confidence level.

D. SVM

SVMs are a relatively new learning process influenced highly by advances in statistical learning theory. This classification divides two separate classes, which are generated from training

Copyright © IJICCT–JUL 2013;Vol 1,Issue 1;ISSN 2347-7202

This journal is cited as : **JIMS 8i-Int'l J. of Inf. Comm. & Computing Technology(IJICCT)**

examples. The overall aim is to generalize well to test data. This is obtained by introducing a separating hyper plane, which must maximize the margin between the two classes; this is known as the optimum separating hyper plane [15].

Applications of SVM:

- SVM is used in Face detection method.
- This is applicable to load forecasting field.
- It also helps in diagnosing cardio-vascular disease.

Advantages of SVM:

- SVM is used in statically learning theory.
- This method is applied for both linear and nonlinear data.
- It uses a nonlinear mapping to transform the original training data into higher dimension.

CONCLUSION

With the rapid expansion in healthcare industry in reference to services and information technology, data mining has become an important part of the healthcare advancement. Nowadays there are various data mining techniques, which are available for application specific area in healthcare industry like in heart disease, to diagnose treatment effectiveness, hospital infection control, clinical data, etc. A comparative study of data mining classification algorithms helps in uncovering the valuable knowledge hidden behind them and in aiding the decision makers to improve the health care services. The presented study of data mining gives medical practitioners and health care planners a tool to help them in quickly comprehending vast clinical databases timely and precisely.

REFERENCES

[1] J. Zvarova and A. Vesely, "Data Mining and Medical Knowledge Management: Cases and Applications", pp. 1-36, 2009. Available at: http://www.igi-global.com/book/data-mining-medical-knowledge-management/231. Last accessed on 14/09/13.

[2] M. Hemalatha and S. Megala, "Mining Techniques in Health Care: A Survey of Immunization", *Journal of Theoretical and Applied Information Technology*, Vol. 25, pp. 63-70, 2011.

[3] K. Srinivas, B. Kavihta Rani and A. Govrdhan , "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", *International Journal on Computer Science and Engineering*, Vol. 2, No. 2, pp. 250-255, 2010.

[4] C.K. Hian and T. Gerald, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management*, Vol. 19, No. 2, pp.64-70, 2005.

[5] K. Kincade, "Data mining: digging for healthcare gold", *Insurance & Technology*, Vol. 23, No. 2, IM2-IM7, 1998.

[6] Milley, "Healthcare and data mining", *Health Management Technology*, Vol. 21, No. 8, pp. 44-47, 2000.

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

[8] J. Chen and R. Greiner, "Comparing Bayesian Network Classifiers", *In Proc. of UAI-99*, pp.101–108, 1999.

[9] "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico. Available at: http://nmhealth.org/ERD/HealthData/Diabetes/SelectedCVDHospitalizationReport2001-2003final.pdf. Last accessed on 14/09/13.

[10] S. Chaudury, and U. Dayal, "An overview of data warehousing and OLAP technology", *ACM SIGMOD Record*, Vol.26, No.1, pp:65-74, 1997.

[11] R. Gaynes, C. Richards, J Edwards, et al., "Feeding back surveillance data to prevent hospital-acquired infections", *Emerg Infect Dis, Vol.* 7, pp.295-298, 2001.

[12] S.E. Brosette, A.P. Spragre, W.T. Jones, S.A. Moser. "A data mining system for infection control surveillance", *Methods Inf Med*, Vol. 39, pp. 303-310, 2000.

[13] P. Cerrito, "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*, San Antonio, TX, November 9-13, 2002.

[14] M. Ridinger, "American Healthways uses SAS to improve patient care", *DM Review, Vol. 12, pp: 139,* 2002.

[15] V. Kumar and L. Verma, "Binary Classifiers for Health Care Databases: A Comparative Study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer", *International Journal of Computer Science and Technology*, Vol.1, pp.124-126, 2010.

[16] M. Kumari and S. Godara , "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction" *International journal of Computer Science and Technology,* Vol. 2 ,No. 2, June 2011.

Copyright © IJICCT–JUL 2013;Vol 1,Issue 1;ISSN 2347-7202

This journal is cited as : **JIMS 8i-Int'l J. of Inf. Comm. & Computing Technology(IJICCT)**