MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Applications of Data Mining in Healthcare

Richi Nayak

School of Information Systems, Queensland University of Technology, Brisbane, Australia

Abstract

In this paper we discuss the use of data mining in many areas of Healthcare such as the treatment effectiveness, healthcare management, customer relationship management and the fraud and abuse detection.

Keywords:

data mining, healthcare informatics

Introduction

Healthcare is one of the world fastest growing industries and profession [7]. One of the main objectives of a healthcare organisation is its ability to obtain, store, analyse and be able to utilise data across the whole organisation effectively. The growth of databases in healthcare practices and researches are evident in the large number of claims databases, registries, electronic medical record databases, disease surveillance systems, and ad hoc research database systems. Pattern-identification tasks such as detecting associations between certain risk factors and outcomes, ascertaining trends in healthcare utilisation, or discovering new models of disease in certain population rapidly becoming striking. As Healthcares emphasise on high level decision making from the collected current and historical data for identifying useful trends and creating summaries of the data, query languages written for databases alone are not able to handle that. In order to effectively use the available data, advanced data analysis techniques such as data mining are required. Data mining (DM), also known as knowledge discovery in databases, has been successfully deployed to find trends and patterns of interest in many domains [8]. This paper explores applications of data mining to Healthcare in areas such as the evaluation of treatment effectiveness, management of healthcare, customer relationship management and the detection of fraud and abuse.

Data mining in clinical care and customer services areas

Data mining is increasingly being used to determine the diagnosis and the most suitable treatments that should be prescribed. It assists (1) to study what factors may interfere with a drug, for example high-fiber foods can interfere with the absorption of certain antidepressants and heart medicines [3], (2) to examine between allergies and a disease to find most effective treatment [10], (3) to correlate the presence of one set of symptoms with another set of symptoms in a diagnosis by analyzing the outcome [12], (4) to find common symptoms to aid diagnosis of an outbreak so that prevention can occur before the outbreaks, and (5) to identify chronic diseases such as who are the high risk patients and how to reduce their number of hospital stay and claims [5]. Data mining techniques can establish the behaviour patterns relevant to a disease. This helps administrators to determine the optimal lengths of hospital stay for specific diseases and to establish best practices for treating diseases [6].

Predictive modeling is immensely useful in predicting the cause of a disease and determining effective treatments based on many known inputs. For example to analyze what symptoms are responsible for thrombosis attacks (blood clots in the heart), a data mining algorithm examines the conditions pre and post the attacks of many patients. Information like: changes in lab results over time, patient demographics, thrombosis diagnosis symptoms leading to thrombosis and other collagen diseases; are extracted from clinical databases. The extracted patterns such as “high APPT levels are 80% responsible for thrombosis” help in interpretation of the patient profile. The accuracy for these kinds of predictions is reported as being above eighty percent [2].

Health Geomatics helps decision makers and service planners to understand the relationship between location and healthcare services [1]. Data mining can be used to evaluate the geography of diseases and healthcare resource utilization. The geography of diseases covers exploration, modeling spatiotemporal incidences of diseases and generating hypotheses of new disease. It helps to identify diseases that occur in a particular area. Health care institutions can then use this information to implement preventative schemes to reduce an outbreak. The information provided to health care officials is more powerful than any spreadsheet database table because it helps to discover new knowledge. Additionally, various information can be integrated to aid health related analysis, this includes: population (census and socioeconomic data), environmental and ecological data, topography, hydrology and climate data, health infrastructure (data on mortality, morbidity, disease distribution and health care facilities) and data needed to perform different types of health related analysis.

Data mining is also been used in modelling and analysis of complex data sets such as genome to understand the human genetic structure [4]. The pattern recognition of a cancer cell based on the genes with the use of predictive data mining technique helps physicians to build more precise treatments. Data mining programs can analyse data at a rate and level of complexity that is impossible to do with other methods. For health care professionals, disease progression is an example where sequence based data mining algorithms may be used. Attributes such as how the viruses or cancer cells spread, the different stages, symptoms etc are taken into consideration while analyzing the path of progression. The sequential patterns helped to determine which organ the disease is likely to spread to or confined to an organ according to the endpoint in the sequence generated.

To improve the clinical and management decision support, the health care industry is increasingly using data mining to determine health policies and priority areas, to coordinate care trials and in improving customer relationship management. Many hospitals have been applying DM techniques to the patients' survey collected. The results give hospitals the knowledge on patient requirements so that they can tailor services to suit patients' needs [9]. Predictive DM is also being used to match the clients' need with a suitably qualified staff member in a home-based health care environment by a health care agency where the coordinator had to match manually otherwise after receiving the request [11].

Application of DM techniques has also been used in financial areas of the healthcare organisation such as fraud detection and monitoring of cost and resources. Association rules are used for analysing the relationships among people, departments and organizations in health informatics A Health insurance organization can benefit from applying association rules to analyze their patient claims, in particular with the steady rise in health-care costs and the growing urgency to control these costs. For example, association rules can be used to detect patterns in ordering pathology services and the claims made by the patients. For example if a patient claims "test-B with test-C, there will be a 92.8% chance that a test-A would also be claimed". The Insurance company can use association rules to prevent fraud and inappropriate practices of requesting or providing services which are unreasonable, unnecessary or excessive (e.g., indiscriminate ordering of blood tests).

Conclusion and future trends

Over the past decade, data mining has been increasingly utilised by many major health institutes. Healthcare organisations have come to realisation that data, stored in the form of databases, are not being used to their maximum capability. Data mining methods are required for decision-making thus assuring quality healthcare. Data mining helps healthcare insurers to detect fraud and abuse; healthcare organisations in customer relationship management decisions; physicians to identify effective treatments and best practices; and patients to receive better and more affordable healthcare services.

The future of data mining in the field of healthcare is still an emerging process. Many aspects of data mining need to be introduced. One cannot depend entirely on a data mining output because its incorrect interpretation of the data can be a matter of life and death. The post-processing in which the outcomes are evaluated should be in depth. Alternative measures such as sensitivity, specificity, post-test probability and misclassification cost should also be considered when evaluating the quality of data mining outcome in a medical domain.

Dealing with more comprehensive data such as ultrasound images, temporal data and mixed data types, is a key direction for the future. In particular, when we have large datasets that contain rare events (temporal sequences), it is due to the risk factors that a straightforward application of data mining might not be able to identify frequent patterns of discovery such as low occurrence rate of adverse drug reactions or biomedical fluid flows during respiration.

Incorporating any associated knowledge from multiple locations significantly increases the efficiency of the mining process and the quality of the knowledge discovered. In particular, there is a wealth of information exists on Internet for medical domain. The mining process should take advantage of existing information in real time while making decision. A standard format for communication also needs to be devised such as XML. Research has shown that the better and improved results are achieved when combined the results of many DM algorithms and operations. Currently, each algorithm produces output in its own format. The integration takes extra processing. When the inputs to and outputs from a model uses the same expression language such as XML, the integration process will be easier.

References

[1] Boulos, M. N. K., Roudsari, A. V., & Carson, E. R. (2001). Health geomatics: an enabling suite of technologies in health and healthcare. *Computers and Biomedical Research, 34*(3), 195 - 219.

[2] Epstein, S., & Denius, H. (2002). Health Care's Rocky Road: Data Mining For Better QualityOf Care Is A Path Worn Well Into The Future. *ADVANCE for Health Information Executives*.

[3] Graedon, J., & Graedon, T. (1996). *The People's Pharmacy*: New York: St. Martin's Griffin.

[4] *Human Genome Project Information.* Retrieved November, 2006, from http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

[5] Kincade, K. (1998). Data mining: Digging for healthcare gold. *Insurance & Technology. New York, 23*(2), IM2.
[6] Kolar, H. R. (2001). Caring for healthcare. *Health Management Technology, 22*(4), 46.
[7] Leeder, S. R. (2003). Achieving equity in the Australian healthcare system. *MJA, 179*(9), 475-478.
[8] Melli, M., Zaiane, O. R., & Kitts, B. (2006). Introduction to the special issue on successful real-world data mining applications. *Explor. Newsl., 8*, 1-2.
[9] Milley, A. (2000). Healthcare and data mining. *Health Management Technology, 21*(8), 44.
[10] Milley, A. (2000). Healthcare and data mining. *Health Management Technology, 21*(8), 44-56.
[11] Nayak, R., & Warren, D. (2002). *Employing Data Mining in a Health Agency System.* The 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Nevada, USA.
[12] Sintchenko, V., & Coiera, E. (2002). *Which clinical decisions benefit from automation? A task complexity approach.* MIE2002: Health Data in the Information Society, Amsterdam.

Address for correspondence

Dr Richi Nayak
School of Information Systems, FIT, QUT
GPO Box 2434, Brisbane, QLD 4001, Australia
Email: r.nayak@qut.edu.au
Phone: +61 7 3138 1976
Fax: +61 7 3138 1214