A Data Mining Framework For Building Health Care Management System

R. Anand [1] , Dr. S. K. Srivatsa [2]

[1]Research Scholar, Sri Chandra Sekarendra Viswa Maha Vidyalaya, Enathur, Kanchipuram-531 602.
[2]Sr.Professor, St. Joseph College of Engineering, Chennai-600 119.

Abstract

Data mining is a relatively new field of research whose major objective is to acquire knowledge from large amounts of data. In medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available. On the one hand, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules. A major objective of this paper is to evaluate data mining tools in medical and health care applications to develop a software with using that tools that can help make timely and accurate decisions. As there are a number of data mining algorithms and tools available we consider only a few tools to evaluate on these applications and develop classification rules that can be used in prediction. In this paper we construct Health Care Information System Software which enables to store all patient's historical data and it can be helpful to medical practitioners to cure the diseases. Health Care Information System (HCIS) provide an effective way to solve the problem of managing clinical data. HCIS has in fact been playing a minor role in the industry for many years but has yet to be implemented successfully end-to-end because of the many hurdles it has faced such as privacy concerns, cost and simply the lack of technology. In this paper we discuss how the health care industry solving their problems through data mining techniques. We discuss learning methods in data mining, data mining tasks, scope of data mining, importance of data mining in health care industry, forecasts and issues in health care industry.

Keywords : Data Mining, Health care, Clustering, K-Nearest Neighbour, Predictions

1 Introduction

As the health care delivery system adopts information technology, vast quantities of health care data become available to mine for valuable knowledge. Health care organizations generally adopt information technology to reduce costs as well as improve efficiency and quality. Medical researchers hope to exploit clinical data to discover knowledge lying implicitly in individual patient health records. These new uses of clinical data potentially affect healthcare because the patient-physician relationship depends on very high levels of trust. To operate effectively physicians need complete and accurate information about the patient. However, if patients do not trust the physician or the organization to protect the confidentiality of their health care information, they will likely withhold or ask the physician not to record sensitive information. This puts the patient at risk for receiving less than optimum care, the organization at risk of having incomplete information for clinical outcome and operational efficiency analysis, and may deprive researchers of important data.

Data mining especially when it draws information from multiple sources poses special problems. For example, hospitals and physicians are commonly required to report certain information for a variety of purposes from census to public health to finance. This often includes patient number, Postal code, sex, date of birth, age, service date, diagnoses codes , procedure codes (CPT), as well as physician identification number, physician postal code, and total charges. Compilations of this data have been released to industry and researchers.

In this paper we explore issues in managing privacy and security of healthcare information used to mine data by reviewing their fundamentals, components and principles as well as relevant laws and regulations. We also present a literature review on technical issues in privacy assurance and a case study illustrating some potential pitfalls in data mining of individually identifiable information. We close the chapter with recommendations for privacy and security good practices for medical data miners.

2 Learning methods in Data mining

As we said before data mining is one among the most important steps in the knowledge discovery process. It can be considered the heart of the KDD process. This is the area, which deals with the application of intelligent algorithms to get useful patterns from the data. Some of the different methods of learning used in data mining and as follows :

Classification learning: The learning algorithms take a set of classified examples (training set) and use it for training the algorithms. With the trained algorithms, classification of the test data takes place based on the patterns and rules extracted from the training set. Classification can also be termed as predicting a distinct class.
Numeric predication: This is a variant of classification learning with the exception that instead of predicting the discrete class the outcome is a numeric value.
Association learning: The association and patterns between the various attributes are extracted are from these rules are created. The rules and patterns are used predicting the categories or classification of the test data.
Clustering: The grouping of similar instances in to clusters takes place. The challenges or drawbacks considering this type of machine learning is that we have to first identify clusters and assign new instances to these clusters.

There are several learning methods that can be used within each type of learning methods (E.g. Decision Tree can be considered as a classification technique, Kth Nearest Neighbor is considered as a clustering technique) but regardless of the learning methods, concept is given to the notation on what is to be learned and concept description is the outcome produced by the instance after the learning procedure.

3 Building Health Care Information System

Healthcare is a very research intensive field and the largest consumer of public funds. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no longer ignore these emerging tools. This resulted in uniting of healthcare and computing to form health care information system. This is expected to create more efficiency

and effectiveness in the health care system, while at the same time, improve the quality of health care and lower cost.

Health informatics is an emerging field. It is especially important as it deals with collection, organization, storage of health related data. With the growing number of patient and health care requirements, having an automated system will be better in organizing, retrieving and classifying of medical data. Physicians can input the patient data through electronic health forms and can run a decision support system on the data input to have an opinion about the patient's health and the care required. An example in the advances in health informatics can be the diagnosis of a patient is health by a doctor practicing in another part of the world. Thus healthcare organizations can share information regarding a patient which will cut costs for communication and at the same time be more efficient in providing care to the patient.

There are other issues like data security and privacy, which is equally important when considering health related data. Thus HCIS "deals with biomedical information, data, and knowledge--their storage, retrieval, and optimal use for problem solving and decision making". This is a highly interdisciplinary subject where fields in medicine, engineering, statistics, computer science and many more come together to form a single field. With the help of smart algorithms and machine intelligence we can provide the quality of healthcare by having, problem solving and decision-making systems. Information systems can help in supporting clinical care in addition to helping administrative tasks. Thus the physicians will have more time to spend with the patients rather than filling up manual forms.

4 The Data Mining Task

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

Exploratory Data Analysis
In the repositories vast amount of information's are available .This data mining task will serve the two purposes
(i)With out the knowledge for what the customer is searching, then
(ii) It analyze the data
These techniques are interactive and visual to the customer.

Descriptive Modeling

It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

Predictive Modeling

This model permits the value of one variable to be predicted from the known values of other variables.

Discovering Patterns and Rules

This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and many more techniques in the data mining algorithm like K-Means. These are called the clustering algorithm.

Retrieval by Content

The primary objective of this task is to find the data sets of frequently used in the for audio/video as well as images It is finding pattern similar to the pattern of interest in the data set

5 The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automated prediction of trends and behaviors.

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Artificial neural networks

Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

Genetic algorithms

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest neighbor method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.

Rule induction

The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

6 The Importance and Uses of Data Mining in Health Care

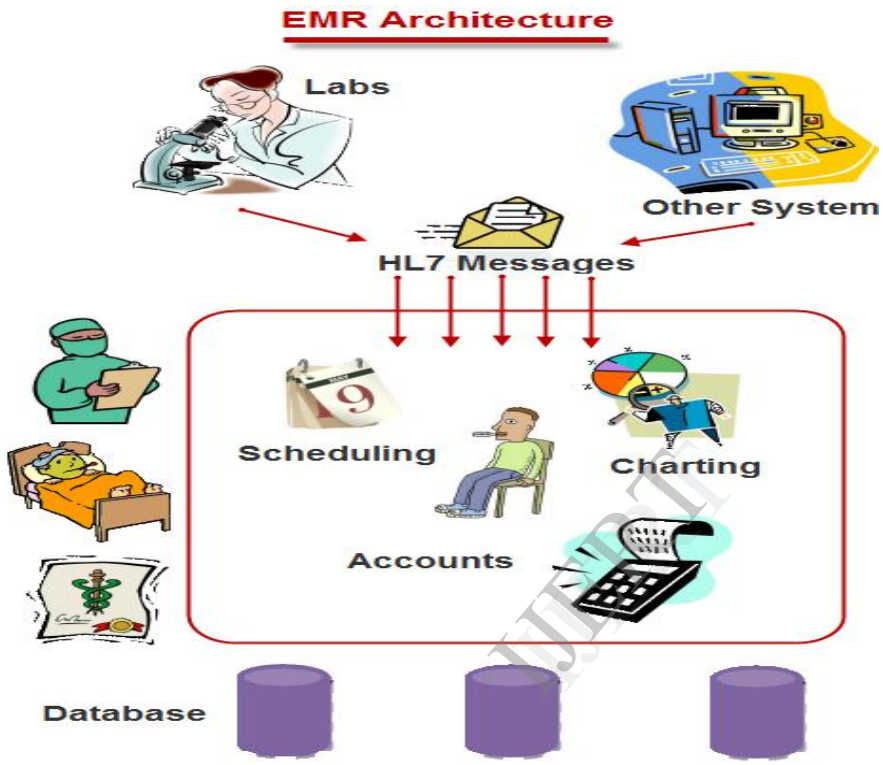
Despite the differences and clashes in approaches, the health sector has more need for data mining today. There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just concerns of public health but also the private health sector (which, in fact, as can be shown later, are also stakeholders in public health).

Data overload. There is a wealth of knowledge to be gained from computerized health records. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge. In fact, some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information. Computers and data mining are best-suited for this purpose.

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 India could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors . By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

Policy-making in public health. Lavrac et al. (2007) combined GIS and data mining using among others, Weka with J48 (free, open source, Java-based data mining tools), to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making."

Applying data mining in the medical field is a very challenging undertaking due to the idiosyncracies of the medical profession. Shillabeer and Roddick's work (2007) cite several inherent conflicts between the traditional methodologies of data mining approaches and medicine. In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practise, which simply starts with the data set without an apparent hypothesis.



EMR – Electronic medical Records used in Health Care information system

Also, whereas traditional data mining is concerned about patterns and trends in data sets, data mining in medicine is more interested in the minority that do not conform to the patterns and trends. What heightens this difference in approach is the fact that most standard data mining is concerned mostly with describing but not explaining the patterns and trends. In contrast, medicine needs those explanations because a slight difference could change the balance between life or death.

For example, anthrax and influenza share the same symptoms of respiratory problems. Lowering the threshold signal in a data mining experiment may either raise an anthrax alarm when there is only a flu outbreak. The converse is even more fatal: a perceived flu outbreak turns out to be an anthrax epidemic (Wong et al 2005). It is no coincidence that we found that, in most of the data mining papers on disease and treatment, the conclusions were almost-always vague and cautious.

7 Forecasting in Health Sector

In general predictions about future health - of individuals and populations - can be notoriously uncertain. However all projections of health care in India must in the end rest on the overall changes in its political economy - on progress made in poverty mitigation (health care to the poor) in reduction of inequalities in generation of employment /income streams (to facilitate capacity to pay and to accept individual responsibility for one's health in public information and development communication (to promote preventive self care and risk reduction by conducive life styles) and in personal life style changes (often directly resulting from social changes and global influences). Of course it will also depend on progress in reducing mortality and the likely disease load, efficient and fair delivery and financing systems in private and public sectors and attention to vulnerable sections- family planning and nutritional services and women's empowerment and the confirmed interest of ensure just health care to the Largest extent possible. To list them is to recall that Indian planning had at its best attempted to capture this synergistic approach within a democratic structure. It is another matter that it is now remembered only for its mixed success.

Available health forecasts

There is a forecast on the new health challenges likely to emerge in India over the next few decades. Murry and Lopez have provided a possible scenario of the burden of disease (BOD) for India in the year 2020, based on a statistical model calculating the change in DALYS are applied to the population projections for 2020 and conversely. The key conclusions must be understood keeping in the mind the tact that the concept of DALYs incorporates not only mortality but disability viewed in terms of healthy years of life lost. In this forecast, DALYs are expected to dramatically decrease in respect of diarrhoeal diseases and respiratory infections and less dramatically for maternal conditions. TB is expected to plateau by 2000, and HIV infections are expected to rise significantly up to 2010. Injuries may increase less significantly, the proportion of people above 65 will increase and as a result the burden of non-communicable disease will rise. Finally cardiovascular diseases resulting any from the risk associated with smoking urban stress and improper diet are expected to increase dramatically.

Under the same BOD methodology another view is available from a four - state analysis done in 1996 these four states - AP, Kamataka, W. Bengal and Punjab - represent different stages in the Indian health transition. The analysis reveals that the poorer and more populated states. West Bengal, will still face a large incidence of communicable diseases. More prosperous states, such as Punjab further along the health transiting will witness sharply increasing incidence of non-communicable diseases especially, in urban areas. The projections highlight that we still operating on unreliable or incomplete base data on mortality and causes of death in the absence of vital registration statistics and know as yet little about how they differ between social classes and regions or about the dynamic patterns of change at work. It also highlights the policy dilemma of how to balance between the articulate middle upper class demand for more access to technologically advanced and subsidized clinical services and the more pressing needs of the poor for coverage of basic disease control interventions. This conflict over deployment of public resources will only get exacerbated in future. What matters most in such estimates are not

societal averages with respect to health but sound data illumining specifically the health conditions of the disadvantaged in local areas that long tradition of health sector analysis looking at unequal access, income poverty and unjustly distributed resources as the trigger to meet health needs of the poor. That tradition has been totally replaced by the currently dominant school of international thought about health which is concerned primarily with efficiency of systems measured by cost effectiveness criteria.

8 Conclusion and Future Work

In this paper we briefly reviewed the various data mining applications in clinical data. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining. Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation. Most of the domain specific data mining applications show accuracy above 90%. The generic data mining applications are having the limitations. From the study of various data mining applications it is observed that, no application called generic application is 100 % generic. The intelligent interfaces and intelligent agents up to some extent make the application generic but have limitations. The domain experts play important role in the different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data mining, and the context parameters. The domain specific applications are aimed to extract specific knowledge. The domain experts by considering the user's requirements and other context parameters guide the system. The results yield from the domain specific applications are more accurate and useful. Therefore it is conclude that the domain specific applications are more specific for data mining. From above study it seems very difficult to design and develop a data mining system, which can work dynamically for any domain. Cellular phones plays a very important role in today's information retrieval system. Some of the new handheld devices, cellular phones, PDAs, the Blackberry and others can be connected to the Internet and information can be received and sent from servers. There are a number of different data mining algorithms that produce rules that can be stored in mobile devices and used for data classification. A possibility for future work could be to implement a local interface for the device where user can input data directly into their mobile devices, and based on the rule set, can deliver the answer back, i.e. classification is done using rules stored in the database of the PDA. This can be a handy tool for medical practitioners. Finally we conclude that our software Health care information system is to solve all clinical data problems.

References

[1] Audain, C. (2007). Florence Nightingale. Online: http://www.scottlan.edu/lriddle/women/ nitegale.htm. Accessed 30 July 2009.

[2] Ayres, I (2008). Super Crunchers. New York: Bantam Books.

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

[4] 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

[5] 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).

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

[7] "Electronic Medical Records, Electronic Health Records, " Oct 12, 2005, http://www.openclinical.org/emr.html (accessed March 3, 2010).

[8] Trisha Torrey,"The Benefits of Electronic Medical Records (EMRs)," Feb 08, 2008, http://patients.about.com/od/electronicpatientrecords/a/EMRbenefits.htm (accessed March 20, 2010).

[9] David York, "What is an EMR - Electronic Medical Record?," Jan 08, 2008, http://www.disabledworld. com/artman/publish/emr.shtml (accessed March 3, 2010).

[10] Shruti Sharma, "Electronic Medical Records Concepts and Data Management", Feb.2011

[11] Islan, M.Z., and Brankovic, L., A. (2004). "Framework for Privacy Preserving Classification in Data Mining, School of Electrical Engineering and Computer Science," *Australasian Computer Science Week.*

[12] Levin, E.G., Arango, J., Steimle, A.E., Lee, P.C., Fireman, B. (2001). "Innovative Approach to Guidelines Implementation Is Associated with Declining Cardiovascular Mortality in a Population of Three Million [abstract]," in *American Heart Association's Scientific Se*ssions, Anaheim, California

[13] Ted Cooper and Jeff Collman, "Managing Information Security and Privacy in Health care data mining"

[14] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[15] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[16] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[17] Hirdes JP, Marhaba M, Smith, TF et al. 2001 Development of the Resident Assessment Instrument - Mental Health (RAI-MH), Hospital Quarterly, 4(2), 44-51

[18] Arun George Eapen (2004), "Application of Data mining in Medical Applications"

[19] Witten, T.H and Frank, E. 2000 Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco.

[20] Dzeroski S, Hristovski D, Peterlin B. Using data mining and OLAP to discover patterns in a database of patients with Y chromosome deletions. *Proc AMIA Symp.* 2000;215–219.

[21] Silver M, Sakata T, Su HC, Herman C, Dolins SB, O'Shea MJ. Case study: how to apply data mining techniques in a healthcare data warehouse. *Healthc Inf Manag.* 2001;15:155–164.

[22] Hristovski D, Rogac M, Markota M. Using data warehousing and OLAP in public health care. *Proc AMIA Symp.* 2000;369–373.