International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN 2249-6831
Vol. 3, Issue 1, Mar 2013, 239-248

A STUDY ON CLINICAL PREDICTION USING DATA MINING TECHNIQUES

V. KRISHNAIAH[1], G. NARSIMHA[2] & N. SUBHASH CHANDRA[3]

[1]Assistant Professor, Department of CSE, CVR College of Engineering, Vastunagar, Hyderabad, Andhra Pradesh, India

[2]Associate Professor, Department of CSE, JNTUH College of Engineering, Kondagattu, Andhra Pradesh, India

[3]Professor of CSE & Principal, Holy Mary Institute of Technology and Science, Hyderabad, Andhra Pradesh, India

ABSTRACT

This paper can present an overview of the applications of data mining techniques, medical, research, and educational aspects of Clinical Predictions. 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 techniques in clinical and health care applications to develop a accurate decisions. The paper also provides a detailed discussion of medical data mining techniques can improve various aspects of Clinical Predictions.

KEYWORDS: Decision Making, Medical Records, Data Mining, Decision Tree, Naive Baye, Association Rule, Outpatient Clinic

INTRODUCTION

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [1]. The ability to use these data to extract useful information for quality healthcare is crucial.

Clinical Prediction is a rapidly growing field that is concerned with applying Computer Science and Information Technology to medical and health data. With the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of Clinical Diagnosis to save time, money, and human lives.

"Clinical Prediction is the computerization of medical information to support and optimize (1) administration of health services; (2) clinical care; (3) medical research; and (4) training. It is the application of computing and communication technologies to optimize health information processing by collection, storage, effective retrieval (in due time and place),analysis and decision support for administrators, clinicians, researchers, and educators of medicine."

Computer assisted information retrieval may help support quality decision making and to avoid human error. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. Imagine a doctor who has to examine 5 patient records; he or she will go through them with ease.

But if the number of records increases from 5 to 50 with a time constraint, it is almost certain that the accuracy with which the doctor delivers the results will not be as high as the ones obtained when he had only five records to be analyzed. Typical problems that data mining addresses are how to classify data, cluster data, find associations between

data items, and perform time series analysis. Numerous data mining techniques have been invented for each type of problem.[2],[3].Each problem requires data mining techniques to analyze large quantities of data.

In this survey, presented an overview of the applications of data mining techniques in various subfields of Clinical Predictions. For each subfield of Clinical Predictions, and also presented how clinical data warehousing in combination with data mining can help administrative, clinical, research and educational aspects of Clinical Predictions. Finally, we discuss a number of unique challenges of data mining in Clinical Predictions.

DATA MINING REVIEW

Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [4]. Fayyad defines data mining as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" [5]. Giudici defines it as "a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database" [6].

Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised) [7].

Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [8] Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [9].

Decision Tree algorithms include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [10]. CART uses Gini index to measure the impurity of a partition or set of training tuples [8]. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data.

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods [11]. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

Association Rule: The central task of association rule mining is to find sets of binary variables that co-occur together frequently in a transaction database, while the goal of feature selection problem is to identify groups of that are strongly correlated with each other with a specific target variable. Association rule has the several algorithms like: Apriori, CDA, DDA, interestingness measure etc.

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule can be 'if a customer buys a dozen eggs, he is 80% likely to also purchase milk.' An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. The association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to

identify the most important relationships. The support is an indication of how frequently the data items appear in the database. The confidence indicates the number of times the if/then statements have been found to be true.

Types of Association Rule Techniques are as Follows

- Multilevel Association Rule
- Multidimensional Association Rule
- Quantitative Association Rule

Prediction: The prediction as it name implied is one of the data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in blood donors to predict the behaviour for the future if we consider donor is an independent variable, blood could be a dependent variable. Then based on the historical data, we can draw a fitted regression curve that is used for donor's behaviour prediction. Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction.

DATA ANALYSIS TASKS IN DATA MINING TECHNIQUES

Several data mining problem types or analysis tasks are typically encountered during a data mining project. Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result. The data mining analysis tasks typically fall into the general categories listed below. For each data analysis task, an example of a useful data analysis technique is presented.

Table 1 is a matrix that summarizes the data mining analysis tasks and the techniques useful for performing these tasks. The table is representative of the many possibilities since the permutations and combinations of data analysis tasks and techniques are numerous.

Data Summarization gives the user an overview of the structure of the data and is generally carried out in the early stages of a project. This type of initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information. Simple descriptive statistical and visualization techniques generally apply.

Segmentation separates the data into interesting and meaningful sub-groups or classes. In this case, the analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization. Automatic clustering techniques can detect previously unsuspected and hidden structures in data that allow segmentation. Clustering techniques, visualization and neural nets generally apply.

Classification assumes that a set of objects—characterized by some attributes or features—belong to different classes. The class label is a discrete qualitative identifier; for example, large, medium, or small. The objective is to build classification models that assign the correct class to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling. Discriminant analysis, decision tree, rule induction methods, and genetic algorithms generally apply.

Prediction is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for

unseen objects; this problem type is also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets generally apply.

Table 1: Data Analysis Tasks and Techniques

DATA ANALYSIS TECHNIQUES	Data Summarization	Segmentation	Classification	Prediction	Dependency Analysis
Correlation Analysis					✓
Decision Trees			✓	✓	
Association Rules					✓

Dependency Analysis deals with finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of an item given information on other data items. Dependency analysis has close connections with classification and prediction because the dependencies are implicitly used for the formulation of predictive models. Correlation analysis, regression analysis, association rules, case-based reasoning and visualization techniques generally apply.

An Overview of Clinical Predictions and Applications in Data Mining Techniques

As mentioned in the introduction, Clinical Predictions can be divided into four main subfields:

- Administration of health services
- Clinical care
- Medical research
- Training.

The following subsections present an overview of each subfield of health Informatics, and how data mining is, or can be, applied to extend and improve each subfield.

Administration of Health Services

Administrators of health care organizations make hundreds of critical decisions on daily basis. As in any administrative position, the quality of these decisions directly depends on the quality of the information that the decisions are based on. For example, the administrators in a hospital need to decide on the amount of supplies and number of staff and free beds required for an upcoming month. To make this decision, the administrators require an accurate prediction of the number of patients to expect during the coming month, and an approximation of how long each patient will remain in the hospital. As another example, the federal and provincial health administrators need to decide whether a disease outbreak is in progress, and if so, what preventive measures will be most effective against it. To make these decisions, the administration requires a system that can accurately predict a disease outbreak, and also model the cost and benefit of different preventive measures.

Clinical Care

Physicians and nurse practitioners make diagnostic decisions and treatment recommendations based on history, medical imaging, lab results and other text or multimedia records of patients. Clinical Predictions allows doctors to have faster access to more relevant information, and thus make more optimal decisions. For instance, a centralized patient record database will allow a physician in a local clinic to have access to all the relevant medical records of the patient, anywhere

in the country. Furthermore, applying data mining techniques on the centralized database will give doctors analytical and predictive tools that go beyond what is apparent from the surface of the data. For instance, a new practitioner can query for all the decisions that previous practitioners have made on a similar case. Similarly, a predictive model can advise doctors whether a certain case would be better treated as an outpatient or an inpatient.

Clinical Decision Support Systems

The applications of Clinical Predictions in health care decision-making are known as (Computer based) Clinical Decision Support System (CDSS) Shortliffe defines a decision support system as "any computer program that is designed to help health professionals to make clinical decisions" [12,13]. Applications of Clinical Decision Support Systems can be categorized into:

***Information Retrieval*:** CDDS can offer search capabilities for medical queries. For instance the "antibiotic assistant" of Health Evaluation through Logic Processing (HELP) system allow doctors to query the hospital experience with previous infections through the last five years [14].

Alerting Systems: A useful application of CDSS is to monitor inputs and check them for predetermined triggers [15]. These alert systems can be simple, like predefined drug-drug or drug allergy conflicts, or complex, such as alerts based on analysis of various lab results and comparison with expected result protocols.

***Reminders*:** unlike alerts that are triggered by a specific change in input data, reminders are triggered by passage of time and are used for periodic tasks such as immunization or diabetes tests [15].

Suggestion Systems: Unlike alerts, which indicate predetermined conditions in input data, suggestion systems are interactive processes that suggest action oriented messages based on their medical knowledge base.

Prediction Models: CDSS prediction models can be categorized into diagnosis (defined as "aiding in the determination of the existence or nature of a disease" [16] and prognosis (defined as “the forecast of the probable outcome of an illness" [16]) [15]. An example of a diagnosis predictor is a model that detects *nosocomial* clinical predictions based on information from Microbiology laboratory, nurse charting, and other sources.

A Study on HELP System

Health Evaluation through Logic Processing (HELP) system is an example of a Clinical Decision Support System that includes alerting systems, suggestion systems, and prediction models [14]. An example of an alerting system used in HELP is a model that monitors patient laboratory results, and has simple rule-based triggered to detect anomalies. A suggestion system included in HELP is a set of computerized protocols for managing care of Adult Respiratory Distress Syndrome (ARDS) patients. Both alerting and suggestion systems in HELP are rule-based models, developed by physicians, nurses, and specialists in medical informatics.

HELP includes two types of prediction models. One of these models is rule-based models, such as the one used in the Adverse Drug Events (ADE) detection system. The ADE detection system predicts the possibility of a drug reaction based on patient history and a set of predefined protocols. Aside from rule-based models, some prediction models in HELP use logistic regression, *e.g.* the model that predicts nosocomial hospital infections based on a number of risk factors.

Medical Research

Most current successful applications of data mining in Clinical Predictions are in the subfield of medical research. The reason is that most of the current health related data are stored in small datasets scattered through various clinics,

hospitals, and research centers. However, most applications of data mining in clinical and administrative decision support systems require homogeneous and centralized data warehouses. On the other hand, data mining techniques can still be successfully applied on small and scattered datasets, and help researchers extract insightful patterns, cause and effect relationships, and predictive scoring systems from currently available data.

The following subsections introduce a number of examples of data mining techniques applied on small datasets for medical research.

A Study on Drug Exposure Side Effects from Mining Pregnancy Data

Chen et al. investigate the possible effects of multiple drug exposures at different stages of pregnancy on preterm birth, using Smart Rule, a data mining technique for generating associative rules [17]. In this work, two subsets of Danish National Birth Cohort (DNBC) dataset are used. The first subset contains 4454 records including 1000 women who were depressed and/or exposed to various active drugs. This set is used for finding the side effects of anti-depression drugs. The second subset contains 6231 records, including 414 preterm cases. This set is used for finding side effects of multiple types of drugs. The authors develop a tree hierarchical model for organizing the generated rules, in order to ease the recognition of interesting rules by human experts. Using this system, the authors claim that they are able to find novel and interesting rules.

A Study on Association Rules and Decision Trees for Disease Prediction

Ordonez applies different classifiers, associative classifier and decision trees, for predicting the percentage of vessel narrowing (LDA, RCA, LCX and LM) compare to a healthy artery [18]. The dataset contains 655 patient records with 25 medical attributes. Three main issues about mining associative rules in medical datasets are mentioned in this work. A significant fraction of association rules are irrelevant and most relevant rules with high quality metrics appear only at low support. On the other hand, the number of discovered rules becomes extremely large at low support.

Hence, association rules are used with constraints. Each item corresponds to the presence or absence of one categorical value or one numeric interval. First constraint is that there is a limit on the maximum item-set size. Second, the items are grouped and in each association, there is at most one from each group. The third constraint is that each item can only appear in antecedent or consequent. The result from associative classifier is compared with two decision tree algorithms: CN4.5 and CART. The authors demonstrate that associative rules can do better than decision trees for predicting diseased arteries.

Education and Training

The fourth subfield of clinical prediction is related to educating new healthcare professionals and retraining and keeping the current staff up-to-date with recent advances in technology. The education and training subfield of Clinical Prediction can be viewed as an instance of the rapidly growing field of e-learning. An increasing interest in applying data mining techniques to e-learning has emerged in recent years, and some of the early applications show promising results [19].

Data mining techniques can benefit all three groups of people who are in contact with a learning system: students, educators, and administrators [19]. Data mining techniques can monitor the success of students at various learning tasks, and recommend relevant resources, materials, and learning paths to achieve a more successful learning experience. For educators, data mining techniques can provide objective feedback of the structure and the content of a course, discover the learning patterns of the students, and cluster learners into smaller groups that have similar educational habits and needs.

Administrators benefit from data mining techniques by learning about the behavior of their users, so they can optimize the servers, distribute network traffic, and learn about the overall effectiveness of the offered educational programs.

The following case study presents an overview of a relatively new Clinical Prediction of data mining technique to find relevant articles for a particular gene.

A Study on Finding Relevant References to Genes and Proteins in Medline Using a Bayesian Approach

Leonard *et al.* apply a Naive Bayesian approach to find cross-references between the symbol of genes and proteins and Medline articles [20]. The authors extract gene and protein symbols from article titles and abstracts, using a dictionary of gene and protein symbols and a dictionary of English words along with a set of rules. A different set of rules is used to find new gene and protein symbols that are not included in the gene and protein symbol dictionary. After assigning articles to identified genes and proteins, a Bayesian estimated probability (EP) based on word frequency is used to find the relevancy of each assigned article to each gene or protein. Hence, only the relevant articles are chosen for each gene or protein and the result will be a set of relevant references for each gene or protein.

A Study of Clinical Decision Support Systems

Some experts present a broader view of CDSS that it is not limited to the clinical care subfield of Clinical Predictions. Ledbetter and Morgan state that the CDSS capabilities are useful in all phases of the clinical process: (a) assessment, (b) planning, (c) intervention, and (d) evaluation [21]. Table 1, taken from their article, describes the potential applications of CDSS for the cases of a patient-specific focus as well as a population-specific (or aggregation based) focus.

Table 2: Potential Applications of CDSS (Taken from [22])

Clinical Process	*Patient Focus (Point-of-Care) Transactional Analysis*	*Population Focus (Retrospective) Aggregate Analysis*
Assessment	• Risk-factor flags • Clinical group membership • Critical value alerts • Assessment templates • Relevant knowledge-base references • Criteria-based alerts	• Opportunities for improvement • At-risk groups • Insight into disease processes • Understanding of current clinical practice • Community health issues
Plan or Intervention	• Allergy warnings • Drug-to-drug interactions • Drug-to-procedure interactions • Procedure-to-procedure interactions • Standardized order templates • Protocol order sets • Criteria-based orders • Drug cost warnings • Procedure cost warnings • Duplicate drug checks • Duplicate procedure checks • Clinical reminders • Relevant knowledge-base references	• Clinical pathway development • Evidence-based practice guidelines • Protocol development • Care standards development
Evaluation	• Critical value alerts • Criteria-based alerts • Variance tracking • Relevant knowledge-base references	• Outcomes measures • Wellness management • Contract management • Clinical risk adjustment

Court right *et al.* has developed a list of core requirements for CDSS tools and the following comprise the major requirements discussed in their article [22]. The CDSS tools need to:

- Have enhanced networking and distributive features
- Be used at all decision making levels in a organization
- Be used in both real time and retrospective modes

- Enabled predictive capabilities using classical statistics
- Utilize "white box" (openly disclosed but protected) methodologies for prediction and detailed support to provide the kind of accuracy rates required for health care decision making

Note that these requirements are not limited to the clinical care subfield of clinical predictions. In addition to the above, we feel that in the broader view, CDSS tools also need to:

- Apply AI techniques for disease prediction
- Use other techniques such as spatial data mining and spatio-temporal data mining to assist in health care decision-making
- Be able to provide a feedback to the decision makers regarding the efficiency of the system
- Have Graphics and graphing capabilities so as to be able to present the data in several formats such tables, bar charts, pie charts, graphs *etc.*
- Have tighter security, and access controls in order to avoid personal data falling into malicious hands.

In the longer term, it is expected that the clinical data can be used to assess "episodes of risk" [22] wherein CDS systems will help in early identification of risk factors such as diet, exercise, travel, and air and water standards. It is also expected that in the future CDS systems will also help in performance benchmarking, continuing medical education of the clinicians by the use of their own data, identification of best practices, creation and utilization of standard terminology etc. [22].

CONCLUSIONS

This paper provided an overview of applications of data mining techniques in administrative, clinical, research, and educational aspects of Clinical Predictions. This paper established that while the current practical use of data mining in health related problems is limited, there exists a great potential for data mining techniques to improve various aspects of Clinical Predictions. Furthermore, the inevitable rise of clinical data will increase the potential for data mining techniques to improve the quality and decrease the cost of healthcare.

ACKNOWLEDGMENTS

One of us, C Madhusudhan Reddy is grateful Chairman to the CVR College of Engineering for providing the research facilities to carry out the research program.

REFERENCES

1. Huang, H. et al. "Business rule extraction from legacy code", Proceedings of 20th International Conference on Computer Software and Applications, IEEE COMPSAC'96, 1996, pp.162-167 J.
2. Brachman, R., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., and Simoudis, E. "Mining Business Databases." Communications of the ACM, 1996, 39(11), 42–48.
3. Brin, S., Motwani, R., Ullman, J., and Tsur, S. "Dynamic Itemset Counting and Implication Rules for Market Basket Data." Paper presented at the SIGMOD Conference, Tucson, Ariz., 1997.
4. Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

5. Fayyad, U: “Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases”, Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.
6. Giudici, P.: “Applied Data Mining: Statistical Methods for Business and Industry”, New York: John Wiley, 2003.
7. Obenshain, M.K: “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.
8. Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.
9. Charly, K.: “Data Mining for the Enterprise”, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.
10. Ho, T. J.: “Data Mining and Data Warehousing”, Prentice Hall, 2005.
11. Tang, Z. H., MacLennan, J.: “Data Mining with SQL Server 2005”, Indianapolis: Wiley, 2005.
12. Shortliffe EH, "Computer programs to support clinical decision making". JAMIA 258, 1987, 61-66.
13. Nykänen P., "Decision Support Systems from a Health Informatics Perspective". Tampere, 2000.
14. Berner E., "Clinical Decicion Support Systems". Springer Science+Business Media, 2007 .
15. Greens R., "Clinical Decision Support". Elsevier Inc., 2007.
16. Canadian Institute of Health Research, http://www.mshri.on.ca/colorectalcancer/definitions.html, 05/25/2008.
17. Chen Y., Henning Pedersen L., Wesley W. Chu, Olsen J., "Drug Exposure Side Effects from Mining Pregnancy Data". ACM SIGKDD Explorations Newsletter, 2007.
18. Ordonez C., " Comparing association rules and decision trees for disease prediction". Proceedings of the international workshop on Healthcare information and knowledge management, 2006.
19. Romero C., Ventura S., "Educational Data Mining: A survey from 1995 to 2005". Expert Systems With Applications, 2007.
20. Leonard JE, Colombe JB, Levy JL, "Finding relevant references to genes and proteins in Medline using a Bayesian approach". Bioinformatics Vol. 18, no. 11, 2002.
21. Ledbetter C. Morgan, M. “Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse”. JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT, VOL 15; PART 2, pages 119-132, 2001.
22. Courtright C., Crawford R. Klubert D. “Criteria for Developing Clinical Decision Support Systems”. CBMS '01: Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems 2001.
23. Ledbetter C. Morgan, M. “Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse”. JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT, VOL 15; PART 2, pages 119-132, 2001.