Mining Medical Databases Using Graph based Association Rules

Wael Ahmad AlZoubi

***Abstract*—Medical databases have accumulated huge amounts of information about patients and their medical conditions. Relationships and patterns within these data can provide new medical knowledge. Sorry to say that few methodologies have been developed and applied to discover this hidden knowledge. In this paper, the graph bases association rules mining (data mining is the main part of Knowledge Discovery in Databases) is used to search for relationships in a large medical database. The data that was collected on 6549 obstetrical patients were evaluated for factors potentially contributing to preterm birth using exploratory factor analysis. This paper describes the processes involved in mining a medical database including data warehousing, data query and cleaning, and data analysis.**

***Index Term*—Graph, medical database, rule mining.**

I. Introduction

Nowadays, the data accumulated in medical databases are progressively growing up quickly, this makes extracting hidden knowledge from medical database complex and more time consuming. Analyzing these data is critical for medical decision makers and managers. The performance of patient management tasks will be improved by analyzing the medical data [1]. Medical data analysis is highly required for the following reasons: 1) Support of specific knowledge-based problem solving activities through the analysis of patient's raw data collected in monitoring [2], 2) Discovery of new knowledge that can be mined through the analysis of groups of case studies, described by symbolic or numeric descriptors [2].

Because of theses reasons, the usual manual data analysis is not enough especially in case of huge database. The best solution to such problems is using, knowledge discovery in databases (KDD), which has been developed quickly in the last few years. KDD is the process of extracting useful knowledge from large datasets. Data mining is the central step in the KDD process, which deals with the problem of extracting interesting, implicit, and useful relations and patterns in data. The association rule mining [3] is one of the best studied models for pattern discovery in the area of data mining. Extracting knowledge from medical databases can be efficiently done by using association rules. Graph based association rules mining simplify the process of generating frequent itemsets (symptoms in medical data) by reading the database only once to build an association graph among frequent items (most occurring symptoms).

Manuscript received January 4, 2013; revised April 28, 2013.

W. A. AlZoubi is with the University Kebangsaan Malaysia (e-mail: wz@ftsm.ukm.my).

We took as a case study the insensitive data in the central database at King Abdullah University Hospital in Jordan using an extensive medical database of obstetrical patients to identify factors that contribute to delivery outputs.

The purpose of this paper is to illustrate how medical production systems such as the King Abdullah University Hospital (KAUH) Delivery Database can be warehoused and mined for knowledge discovery. The eventual goal of this knowledge discovery effort is to recognize factors that can improve the quality and cost impact of delivery care. A data warehouse is constructed for the computer-based patient record system using the collected medical data

The rest of this paper is organized as following: Section II presents the related work, Section III displays the methodology adopted to complete the proposed work, Section IV shows the results of the proposed work, and Section V concludes the paper.

II. Related Work

Recently, the knowledge mining applications from medical databases has been increased rapidly. There are two classes of mining techniques applied on medical data: explanatory and exploratory [4]. Explanatory mining refers to techniques that are used for the purpose of verification or decision making. Exploratory mining is data investigation usually performed at an early stage of data analysis in which an exact mining objective has not yet been set [5].

In the last few years, the number of studies using different techniques of learning on explanatory mining in medical data has been increased progressively. Genetic programming technique has been applied to find out classification rules from medical data sets [6]. Breast cancer survivability has been worked on using AdaBoost algorithms [7]. The fuzzy modeling idea has also been developed on selected features medical data [8]. A system to extract association rules from health examination data has been proposed, after that a case-based reasoning model is used to support the continual disease analysis and management [9]. A different rule mining method with case-based reasoning has been applied recently [10]. Medical data warehouses have been constructed [11] as an extension to the normal medical databases.

On the other hand, very few studies talk about the exploratory mining techniques to extract rules from medical databases. And so, this paper deals with exploratory mining technique. Knowledge visualization in the study of hepatitis patients is one of the studies that use exploratory rule mining in the field of medical data [12]. Another study goes to improve visualization by using the functionality of OLAP tools [13].

III. Methodology

A. Production System Database

The production system database identified for mining was the computer-based patient record system developed at KAUH over the last 10 years. The data collected include demographics, study results, problems, therapies, allergies, subjective and physical findings, and encounter summaries. The data structure uses a proprietary class-oriented approach which stores all of the patient's information in a single record.

The specific database selected for this project was the delivery database used by the Department of Obstetrics and Gynecology at King Abdullah University Medical Center. This database continues to serve as the repository for a regional delivery computerized patient record that is used in inpatient and outpatient settings [14]. The on-line delivery database contains comprehensive data on over 45,000 unique patients collected over nearly 10 years. Additional patient data from the previous decade is also available on tape archive. This computerized repository contains more than 400 medical variables collected on over 20,000 pregnancies and births from a variety of county areas, making it one of the largest and most comprehensive obstetrical datasets available for analysis in Jordan, according to the medical management of the hospital.

B. Data Warehouse Creation

The data warehouse has been created on a centralized server dedicated to the field of data mining queries. Using a method previously described; the medical data was mapped from the graph data structure into relational tables in the personal computer environment. Microsoft SQL Server Version 4.2 was chosen as the database engine and was installed on a PC server with a 160 GHz Pentium CPU, 170 gegabytes of hard disk, 2048 megabytes of RAM, and using Windows NT Server 3.5 operating system.

In order to extract and clean the dataset for analysis for the purposes of this study, a sample two-year dataset (2006-2007) from the data warehouse has been created to be mined for knowledge discovery. Multiple SQL queries are run on the data warehouse to create the dataset. As each variable is added to the dataset, it is cleansed of erroneous values, data inconsistencies, and formatting discrepancies. This cleaning process was accomplished using Paradox Application Language scripts to selectively identify problems and correct the errors.

The crucial role of these scripts was to scan the dataset and convert alphanumeric fields into numerical variables in order to permit statistical analysis. After checking to see if data values were collected during or pertaining to the preterm course of the infant, the script ensured that multiple values for the same variable were not present. If such values existed, the value that was recorded closest to delivery or conception, depending on perceived data quality for the particular variable, was loaded into the final dataset.

The final script identified missing values and prompted the user to either substitute them with an average value for the variable, or to delete the subject from the dataset. Robust demographic variables such as age, race, education, and marital status are automatically selected for inclusion in the dataset, while other routinely collected data elements are randomly selected for inclusion in the study. These elements originated in the problem section and the subjective and physical findings section of the electronic patient records.

C. Mining the Dataset

For this preliminary study, we select exploratory factor analysis for data mining because it had previously been used successfully to explore claims and financial databases in obstetrics [2].

Factor analysis is a statistical method used to identify which data elements can be combined to explain variations between patient groups. This mining technique is appropriate in research problems in which a large number of subjects are compared on a set of variables for which there is no designation of independence or dependence [3].

The statistical software used to conduct the factor analysis was SPSS version 5.0 for Windows.

IV. Results Analysis

A. Creating the Medical Datawarehouse

Following successful transfer of the entire production system database into the SQL Server data warehouse in the personal computer environment, the warehouse contained 45,922 patient records. These records included 15,626 encounters; 29,610 historical data elements; 25,163 individual lab results; 17,453 problems and procedures; and 16,313 subjective and physical findings.

B. Extracting and Cleansing a Test Dataset

The average speed of the queries directed against the data warehouse was roughly 3 minutes, while the longest query for the study required 12 minutes to complete and occurred against a table of nearly 1 million records. The test dataset extracted from the data warehouse contained data regarding 6549 births occurring between January 1, 2006, and December 31, 2007.

TABLE I: Unusable Data Values Encountered While Extracting and Cleaning the Dataset Variables For Analysis

Reason Unusable	Count	Percent of Total Values
Missing values when required	2,213	33.8%
Incomplete dates	249	3.8%
Redundancy of data	4,071	62.16%
Other errors	16	0.24%
Total	6,549	100%

As is shown in Table I, the data cleaning programs used in creating the dataset revealed that 9.47% of the total values in the database were unusable for the purposes of the factor analysis. Data included out of range values such as invalid

weights, format discrepancies such as a date in a numeric field, and data inconsistencies such as two different heights for the same patient combined in one group. While unusable for analysis, the option to store free-text and incomplete dates represent legal values and are not considered as data errors. Thus, only 35% of the unusable values were actually caused by erroneous data.

Factor analysis was successfully conducted on the extracted dataset from the data warehouse. All analyses used list-wise deletion of cases with missing values, and principal components analysis. Preliminary results identified three latent factors that accounted for 48.9% of the variance in the sample chosen to be examined.

REFERENCES

[1] N. Lavrac, E. T. Keravnou, and B. Zupan, *Intelligent Data Analysis in Medicine and Pharmacology: An Overview*, Dordrecht: Kluwer Academic Publishers, 1996, pp. 1–13.

[2] M. Delgado, D. SaÂnchez, M. J. Martõ Ân-Bautista, and M. Vila, "Mining association rules with improved semantics in medical databases," *Artificial Intelligence in Medicine*, vol. 21, pp. 241–245, 2001.

[3] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. the 1993 ACM SIGMOD Conf.*, 1993, pp. 207–216.

[4] J. Roddick, P. Fule, and W. Graco, "Exploratory medical knowledge discovery: experiences and issues," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 94–99, 2003.

[5] K. Kerdprasop and N. Kerdprasop, "SUT-Miner: a knowledge mining and managing system for medical databases," in *Proc. 20th International Workshop on Database and Expert Systems Application*, 2009, pp. 318–322.

[6] C. Bojarczuk, H. Lopez, A. Freitas, and E. Michalkiewicz, "A constrained-syntax genetic programming system for discovering classification rules: Application to medical data sets," *Artificial Intelligence in Medicine*, pp. 27–48, 2004.

[7] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Breast cancer survivability via AdaBoost algorithms," in *Proc. 2nd Australasian Workshop on Health Data and Knowledge Management*, 2008, pp. 55–64.

[8] S. Ghazavi and T. Liao, "Medical data mining by fuzzy modeling with selected features," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 195–206, 2008.

[9] M. Huang, M. Chen, and S. Lee, "Integrating data mining with case based reasoning for chronic diseases prognosis and diagnosis," *Expert Systems with Applications*, vol. 32, pp. 856–867, 2007.

[10] Z. Zhuang, L. Churilov, and F. Burstein, "Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners," *European J. Operational Research*, vol. 195, no. 3, pp. 662–675, 2009.

[11] T. Sahama and P. Croll, "A data warehouse architecture for clinical data warehousing," in *Proc. 12th Australasian Symposium on ACSW Frontiers*, 2007, pp. 227–232.

[12] D. Nguyen, T. Ho, and S. Kawasaki, "Knowledge visualization in hepatitis study," in *Proc. Asia-Pacific Symposium on Information Visualization*, 2006, pp. 59–62.

[13] S. Palaniappan and C. Ling, "Clinical decision support using OLAP with data mining," *Int. J. Computer Science and Network Security*, vol. 8, no. 9, pp. 290–296, 2008.

[14] F. Elevitch, A. Silvers, and J. Sahl, "Projecting corporate health plan utilization and charges from annual ICD-9-CM diagnostic rates: a value-added opportunity for pathologists," *Arch Pathol Lab Med. Nov*, vol. 121, no. 11, pp. 1187–1191, 1997.

Wael A. AlZoubi was born in Irbid, Jordan in 1978. He is now a Ph.D. candidate at University Kebangsaan Malaysia. He got his master of computer science from Yarmouk University in Jordan in 2004. He is interested in mining association rules from transactional data. He is now a lecturer in computer science department at Balqa Applied University, Jordan.