Chapter 4.13
Challenges in Data Mining on Medical Databases

Fatemeh Hosseinkhah
Howard University Hospital, USA

Hassan Ashktorab
Howard University Hospital, USA

Ranjit Veen
American University, USA

M. Mehdi Owrang O.
American University, USA

INTRODUCTION

Modern electronic health records are designed to capture and render vast quantities of clinical data during the health care process. Technological advancements in the form of computer-based patient records software and personal computer hardware are making the collection of and access to health care data more manageable. However, few tools exist to evaluate and analyze this clinical data after it has been captured and stored. Evaluation of stored clinical data may lead to discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. A common goal of the medical data mining is the detection of some kind of correlation, for example, between genetic features and phenotypes or between medical treatment and reaction of patients (Abidi & Goh, 1998; Li et al., 2005). The characteristics of clinical data, including issues of data availability and complex representation models, can make data mining applications challenging.

BACKGROUND

Knowledge discovery in databases (KDD) is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Adriaans & Zantinge, 1996; Han & Kamber, 2001). Data mining is one

step in the KDD where a discovery-driven data analysis technique is used for identifying patterns and relationships in datasets. Recent advances in medical science have led to revolutionary changes in medical research and technology and the accumulation of a large volume of medical data that demands in-depth analysis. The question becomes how to bridge the two fields, data mining and medical science, for an efficient and successful mining of medical data.

While data analysis and data mining methods have been extensively applied for industrial and business applications, their utilization in medicine and health care is sparse (Abadi & Goh, 1998; Babic, 1999; Brossette, Sprague, Hardin, Jones, & Moser, 1998). In Ohsaki, Yoshinori, Shinya, Hideto, and Takahira (2003), the authors discuss the methods of obtaining medically valuable rules and knowledge on pre- and post-processing and the interaction between system and human expert using the data of medical tests results on chronic hepatitis. They developed the system based on the combination of pattern extraction with clustering and classification with decision tree and generated graph-based rules to predict prognosis. In Tsumoto (2000), the author focuses on the characteristics of medical data and discusses how data miner deals with medical data. In (Ohsaki et. al., 2007), authors discuss the usefulness of the interestingness measures for medical data mining through experiments using clinical datasets on meningitis. Based on the outcomes of these experiments, they discuss how to utilize these measures in postprocessing.

The data mining techniques such as Neural Network, Naïve Bayes, and Association rules are at present not well explored on medical databases. We are in the process of experimenting with a data mining project using gastritis data from Howard University Hospital in Washington, DC to identify factors that contribute to this disease. This project implements a wide spectrum of data mining techniques. The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of patient care.

In this article, we discuss the challenges facing the medical data mining. We present and analyze our experimental results on gastritis database by employing different data mining techniques such as Neural Network, Naive Bayes, and Association rules and using the data mining tool XLMiner (Shmueli, Patel, & Bruce, 2007; XLMiner, 2007).

MEDICAL DATA MINING: CHALLENGES

The application of data mining, knowledge discovery and machine learning techniques to medical and health data is challenging and intriguing (Abidi & Goh, 1998; Brossette et al., 1998; Cios & Moore, 2002). The datasets usually are very large, complex, heterogeneous, and hierarchical and vary in quality. Data preprocessing and transformation are required even before mining and discovery can be applied. Sometimes the characteristics of the data may not be optimal for mining or analytic processing. The challenge here is to convert the data into appropriate form before any leaning or mining can begin.

There are a number of issues that must be addressed before any data mining can occur. In the following, we overview some of the challenges that face the data mining process on medical databases (Tsumoto, 2000).

High Volume of Data

Due to the high volume of the medical databases, current data mining tools may require extraction of a sample from the database (Cios & Moore, 2002; Han & Kamber, 2001). Another scheme is to select some attributes from the database. In both approaches, domain knowledge can be used to eliminate irrelevant records or attributes in reducing the size of the database (Owrang, 2007).