Chapter XXIX
Hybrid Data Mining for Medical Applications

Syed Zahid Hassan
Central Queensland University, Australia

Brijesh Verma
Central Queensland University, Australia

ABSTRACT

This chapter focuses on hybrid data mining algorithms and their use in medical applications. It reviews existing data mining algorithms and presents a novel hybrid data mining approach, which takes advantage of intelligent and statistical modeling of data mining algorithms to extract meaningful patterns from medical data repositories. Various hybrid combinations of data mining algorithms are formulated and tested on a benchmark medical database. The chapter includes the experimental results with existing and new hybrid approaches to demonstrate the superiority of hybrid data mining algorithms over standard algorithms.

INTRODUCTION

In the last few decades, medical disciplines have become increasingly data-intensive. The advances in digital technology have led to an unprecedented growth in the size, complexity, and quantity of collected data, that is, medical reports and associated images. According to Damien McAullay (Damien, Graham, Jie, & Huidong, 2005), "there are 5.7 million hospitals admissions, 210 million doctor's visits, and a similar number of prescribed medicines dispensed in Australia annually" (p. 381). All records are captured electronically. There are billions of healthcare records transaction that occur world wide every year.

On the other hand, patient-centered medical applications (e.g., electronic patient records, personal health record, electronic medical records, etc.) are also on the verge of becoming practical, further increasing data growth and leading to a data-rich but information-poor healthcare system. Thus, it has become crucial for data mining researchers to investigate and propose a novel approach that can appropriately utilize such valuable data to provide useful evidence as a basis for future medical practice. The paramount important factor is to utilize the collected data that suit specific and useful purposes which leads to enable the discovery of new "knowledge" that provides insights to assists

healthcare analyst and policy makers to make strategic decisions and predict future consequences by taking into account the actual outcomes of current operative values.

In addition, the world health organization (Gulbinat, 1997) identifies some possible needs for the discovery of knowledge from medical data repositories; this includes, but is not limited to, medical diagnosis and prognosis, patient health planning and development, healthcare system monitoring and evaluation, health planning and resource allocation, hospital and health services management, epidemiological and clinical research, and disease prevention.

Lately, this abundance of healthcare data has resulted in a large number of concerted efforts to inductively discover "useful" knowledge from the collected data, and indeed interesting results have been reported by many researchers. However, despite the noted efficacy of the knowledge discovery method—known as data mining (DM) algorithm—the challenge facing healthcare practitioners today is about data usability and impact, that is, the use of "appropriate" data mining algorithms with the right data to discover value-added "action-oriented" knowledge in terms of data-mediated decision-support services.

Notably, recent advances in data mining algorithms such as neural networks (NN), statistical modeling, evolutionary algorithms, and visualization tools have made it possible to transform any kind of raw data into high level knowledge. However, the main problem is that each method has its own approach to deal with data structure, shape, and validity. This limitation affects the performance of classification systems. Consequently, the need of a hybrid data mining approach is widely recognized by the data mining community (George, & Derek, 2004, p. 151). The number of hybrid data mining endeavours has been initiated all over the globe. The limitations associated with many existing hybrid approaches are: (1) most of the existing approaches either heavily dependant on intelligent methods or statistical methods, barely ensembled to take the advantage of both computations, that is, intelligent and statistical; and most importantly (2), existing approaches generally do not utilize the data for "secondary purposes," such as organisation planning, decision making, forecasting, outcomes and trending, and so forth.

To this end, we argue that there is a need for a hybrid DM approach which is an effective combination of various DM techniques, in order to utilize the strengths of each individual technique and compensate for each other's weaknesses. The aim of this chapter is to present current state-of-art data mining algorithms and their applications and propose a new hybrid data mining approach for clustering and classification of medical data. This chapter aims to further explore the data mining intelligent and statistical machine learning techniques, including supervised and unsupervised learning techniques as well as some effective conventional techniques and systems commonly used in the medical domain.

REVIEW OF EXISTING DATA MINING ALGORITHMS

In this section we discuss the theoretical and technical aspects of data mining and machine learning techniques, data mining algorithms, hybrid approaches, and their applications in medical domain.

Data Mining and Machine Learning

There is some confusion about the terms data mining and knowledge discovery in databases (KDD). Often these two terms are used interchangeably (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1997, p. 154). The term KDD can be denoted to overall process of turning low-level data into high-level knowledge, whereas data mining can be defined as the extraction of useful patterns from the raw data.

The data mining step usually takes a small part of overall KDD process. More specifically, data