Chapter 5
Classification Techniques and Data Mining Tools Used in Medical Bioinformatics

Satish Kumar David
King Saud University, Saudi Arabia

Amr T. M. Saeb
King Saud University, Saudi Arabia

Mohamed Rafiullah
King Saud University, Saudi Arabia

Khalid Rubeaan
King Saud University, Saudi Arabia

ABSTRACT

Increasing volumes of data with the increased availability information mandates the use of data mining techniques in order to gather useful information from the datasets. In this chapter, data mining techniques are described with a special emphasis on classification techniques as one important supervised learning technique. Bioinformatics tools in the field for medical applications especially in medical microbiology are discussed. This chapter presents WEKA software as a tool of choice to perform classification analysis for different kinds of available data. Uses of WEKA data mining tools for biological applications such as genomic analysis and for medical applications such as diabetes are discussed. Data mining offers novel tools for medical applications for infectious diseases; it can help in identifying the pathogen and analyzing the drug resistance pattern. For non-communicable diseases such as diabetes, it provides excellent data analysis options for analyzing large volumes of data from many clinical studies.

DOI: 10.4018/978-1-5225-7077-6.ch005

INTRODUCTION

Developments in information technology have led to significant advancements in how the large volumes of data are handled. Advances in the healthcare have created enormous medical data in the form of electronic health records. All the medical information and history of patients are stored in the electronic health records. Many countries have even set up unique registries for diseases. With advancements in the biomedical research data from genomics, proteomics and metabolomics have flooded the researchers. Appropriate data analysis is necessary to convert these enormous volumes of raw data into meaningful and valuable results. Medical data analysis can be beneficial in the epidemiology and disease surveillance, to predict the pattern of diseases and track the outbreaks. It can be used to analyze the clinical data to evaluate the effectiveness of health programs and identify the people at risk for developing adverse health outcomes. Medical data along with data from other biomedical research can be useful in the development of a faster, economical and effective new drug discovery and development programs. Therefore, medical data analysis has become an important tool for all the stakeholders involved in the healthcare.

Data analysis requires appropriate tools to be effective. Managing the big data has developed into an important field of research known as data mining. It is a method of discovering information from studying the data of medicine, genetics, bioinformatics and education (Fayyad & Stolorz, 1997). Data mining extracts data patterns in large data sets identifying novel, potentially useful and valid information from the data (Fayyad & Stolorz, 1997). It is an incredible potential tool, which can predict patterns, behaviors and can be actualized on existing programming and hardware platforms. Data mining is bolstered by three innovations, such as massive data accumulation, powerful multiprocessor PCs and data mining algorithms. Data mining methods are not the same as traditional statistical strategies though many processes of data mining can be done using statistical methods. Traditional statistical strategies require a lot of user collaboration with a specific goal to approve the accuracy of a model. Therefore, these strategies can be hard to mechanize. Whereas, data mining strategies are appropriate for expansive data collections and can be automated easily. Data mining includes tasks such as deviation recognition, which identifies irregular data records, dependency demonstration also known as market basket analysis that looks for the association between variables, clustering, classification, regression, and summarization (Figure 1). It utilizes modeling, building a model in one circumstance where you know the appropriate response and afterward apply it to another circumstance. It requires knowledge from large dataset to develop models that can analyze the current data. Moreover, unlike other methods, data mining tools do not modify the data to analyze it.

Data mining has two techniques, namely unsupervised and supervised learning techniques. Unsupervised learning technique analyses the data and creates hypothesis to build a model. It is not guided by the variable. Clustering is one of the commonly used unsupervised technique (Guerra et al., 2011). In case of supervised learning technique, the model is built before the analysis. Classification, Statistical regression and Association rules are the commonly used supervised learning techniques in medical field (Yoo et al., 2012).

Moreover, these techniques are used widely in the field of infectious disease control. These include pathogen identification and typing and comparison with the produced molecular profiles with the pre-existing databases such as Institute Pasteur MLST. The phylogenetic analysis that uses different classification techniques such as neighbor-joining and Bayesian analysis. In addition to pathogenomics, that is mainly dependent on data mining of the huge amount of sequence data generated by next-generation sequencing techniques, as authors will discuss later.