Chapter 4.3
Pattern Discovery from Biological Data

Jesmin Nahar
Central Queensland University, Australia

Kevin S Tickle
Central Queensland University, Australia

A B M Shawkat Ali
Central Queensland University, Australia

ABSTRACT

Extracting useful information from structured and unstructured biological data is crucial in the health industry. Some examples include medical practitioner's need to identify breast cancer patient in the early stage, estimate survival time of a heart disease patient, or recognize uncommon disease characteristics which suddenly appear. Currently there is an explosion in biological data available in the data bases. But information extraction and true open access to data are require time to resolve issues such as ethical clearance. The emergence of novel IT technologies allows health practitioners to facilitate the comprehensive analyses of medical images, genomes, transcriptomes, and proteomes in health and disease. The information that is extracted from such technologies may soon exert a dramatic change in the pace of medical research and impact considerably on the care of patients. The current research will review the existing technologies being used in heart and cancer research. Finally this research will provide some possible solutions to overcome the limitations of existing technologies. In summary the primary objective of this research is to investigate how existing modern machine learning techniques (with their strength and limitations) are being used in the indent of heartbeat related disease and the early detection of cancer in patients. After an extensive literature review these are the objectives chosen: to develop a new approach to find the association between diseases such as high blood pressure, stroke and heartbeat, to propose an improved feature selection

DOI: 10.4018/978-1-60960-818-7.ch4.3

method to analyze huge images and microarray databases for machine learning algorithms in cancer research, to find an automatic distance function selection method for clustering tasks, to discover the most significant risk factors for specific cancers, and to determine the preventive factors for specific cancers that are aligned with the most significant risk factors. Therefore we propose a research plan to attain these objectives within this chapter. The possible solutions of the above objectives are: new heartbeat identification techniques show promising association with the heartbeat patterns and diseases, sensitivity based feature selection methods will be applied to early cancer patient classification, meta learning approaches will be adopted in clustering algorithms to select an automatic distance function, and Apriori algorithm will be applied to discover the significant risks and preventive factors for specific cancers. We expect this research will add significant contributions to the medical professional to enable more accurate diagnosis and better patient care. It will also contribute in other area such as biomedical modeling, medical image analysis and early diseases warning.

INTRODUCTION

More frequently clinical decisions are often made based on medical practitioner knowledge and experience rather than on the knowledge hidden in the huge database. The limitations of this practice include unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients (Palaniappan & Awang, 2008). Therefore, it is important to discover the hidden knowledge from a medical database to provide a better care of patient. Li et al., (2004) argue:

- 'Data mining techniques can be successfully applied to ovarian cancer detection with a reasonably high performance' (Li et al., 2004).
- 'The classification using features selected by the genetic algorithm consistently outperformed those selected by statistical testing in terms of accuracy and robustness' (Li et al., 2004).

Similarly many researchers in the machine learning community found that molecular level classification of human tissues has produced remarkable results, and indicated that the gene expression method could significantly aid in the development of efficient cancer diagnosis and classification platforms (Brown et al., 2000; Lu & Han, 2003; Statnikov et al., 2005; Yeh et al., 2007). Machine learning techniques are also widely explored in heart diseases (Ordonez et al., 2001), computer-based medical picture interpretation method contain a major application area provide significant support in medical diagnosis (Coppini et al., 1995; Zhu & Yan, 1997) and many more.

Magoulas & Prentza (2001) suggest, since the understanding of biological systems is not complete, so there are essential features and information hidden in the physiological signals which are not readily apparent. Moreover, the effects between the different subsystems are not distinguishable. Basically biological signals are characterized by substantial variability, caused either by spontaneous internal mechanisms or by external stimuli. Associations between the different parameters may be too complex to be solved with conventional techniques. Modern Machine Learning (ML) methods rely on these sets of data, which can be produced easier, and can help to model the nonlinear relationships that exist between these data, and extract parameters and features which can improve the current medical care (Magoulas, & Prentza, 2001).

As shown in Figure 1 that cancer and heart disease are the top two causes of death in the United Kingdom. Finding patterns in the data that can assist in the early detection of these diseases will have a significant impact on human health.