Chapter XII

Data Mining in Gene Expression Data Analysis: A Survey

Jilin Han, University of Oklahoma, USA
Le Gruenwald, University of Oklahoma, USA
Tyrrell Conway, University of Oklahoma, USA

Abstract

The study of gene expression levels under defined experimental conditions is an important approach to understand how a living cell works. High-throughput microarray technology is a very powerful tool for simultaneously studying thousands of genes in a single experiment. This revolutionary technology results in an extensive amount of data, which raises an important question: how to extract meaningful biological information from these data? In this chapter, we survey data mining techniques that have been used for clustering, classification and association rules for gene expression.
data analysis. In addition, we provide a comprehensive list of currently available commercial and academic data mining software together with their features. Lastly, we suggest future research directions.

Introduction

Recently, bioinformatics has attracted a lot of attention from biologists and computer scientists. One of the most important aspects of bioinformatics is the application of data mining tools to extract meaningful biological information from gene expression data. The study of gene expression levels at a given time or under established conditions is an important approach to understand how a living cell works (Vingron & Hoheisel, 1999). High-throughput microarray technology (Ramsay, 1998; Harrington, Rosenow, & Retief, 2000; Lipshutz et al., 2000; Jordan, 2001) is a powerful tool for simultaneously studying thousands of genes in a single experiment. This revolutionary technology results in an extensive amount of data, which raises a challenging question: how can meaningful biological information be extracted from these data? Important biological information associated with these data may not be discovered or may be misinterpreted due to lack of appropriate and effective data analysis tools and techniques.

Data mining is one of the most important and difficult tasks in gene expression data analysis. Data mining typically includes clustering, classification, and association rule discovery (Lin & Johnson, 2002; Wei, 2002; Johnson & Wichern, 1998; Mirkin, 1996). With extensive microarray data available, clustering can be used to identify genes that are co-regulated in a similar manner under different experimental conditions. Classification provides a way to identify the differences between tissue types such as between normal cells and cancer cells, which facilitates diagnosis of diseases. Discovery of association rules can help biologists to identify genes that govern the expression of other genes in regulatory pathways.

This chapter is organized as follows. First, a background on biology and microarray technology is presented. Then, we discuss the gene expression data characteristics and presentation. Following are reviews of the existing clustering, classification and association rules mining algorithms that have been applied to gene expression analysis. A comprehensive list of available commercial and open source data mining software with their features is then presented. Lastly, we suggest directions for future work.
Identifying Critical Success Factors for Supply Chain Excellence
[www.igi-global.com/chapter/identifying-critical-success-factors-supply/65971?camid=4v1a](www.igi-global.com/chapter/identifying-critical-success-factors-supply/65971?camid=4v1a)

Decision Support-Related Resource Presence and Availability Awareness for DSS in Pervasive Computing Environments
[www.igi-global.com/article/decision-support-related-resource-presentation/1742?camid=4v1a](www.igi-global.com/article/decision-support-related-resource-presentation/1742?camid=4v1a)