Chapter 2
Microarray Data Mining:
Issues and Prospects

Giulia Bruno
Politecnico di Torino, Italy

Alessandro Fiori
Politecnico di Torino, Italy

ABSTRACT

Microarray technology is a powerful tool to analyze thousands of gene expression values with a single experiment. Due to the huge amount of data, most of recent studies are focused on the analysis and the extraction of useful and interesting information from microarray data. Examples of applications include detecting genes highly correlated to diseases, selecting genes which show a similar behavior under specific conditions, building models to predict the disease outcome based on genetic profiles, and inferring regulatory networks. This chapter presents a review of four popular data mining techniques (i.e., Classification, Feature Selection, Clustering and Association Rule Mining) applied to microarray data. It describes the main characteristics of microarray data in order to understand the critical issues which are introduced by gene expression values analysis. Each technique is analyzed and examples of pertinent literature are reported. Finally, prospects of data mining research on microarray data are provided.

INTRODUCTION

With the developing of new technologies and revolutionary changes in biomedicine and biotechnologies, there was an explosive growth of biological data during the last few years. Genome wide expression analysis with DNA microarray technology has become a fundamental tool in genomic research. Since microarray technology was introduced, scientists started to develop informatics tools for the analysis and information extraction from this kind of data. Due to the characteristics of microarray data (i.e. high levels of noise, high cardinality of genes, small samples size) data mining approaches has become the suitable tools to perform any kind of analysis on these data. Many techniques can be applied to analyze microarray data, which can be grouped in four categories: Classification, Feature Selection, Clustering and Association Rule Mining.

DOI: 10.4018/978-1-60960-067-9.ch002

Classification is a procedure used to predict group membership for data instances. Given a training set of samples with a specific number of attributes (or features) and a class label (e.g., a phenotype characteristic), a model of classes is created. Then, the model is exploited to assign the appropriate class label to new data. Model quality is assessed by means of the classification accuracy measure, i.e., the number of correct label predictions over the total number of unlabeled data. The classification of microarray data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors.

Since genetic data are redundant and noisy, and some of them do not contain useful information for the problem, it is not suitable to apply the classification directly to the whole dataset. Feature Selection techniques are dimensional reduction methods usually applied before classification in order to reduce the number of considered features, by identifying and removing the redundant and useless ones. Moreover, feature selection algorithms applied to microarray data allow identifying genes which are highly correlated with the outcome of diseases. Another way to identify redundant genes is to group together sets of genes which show a similar behavior, and then select only a representative for the group. Furthermore, genes with similar expression pattern under various conditions or time course may imply co-regulations or relations in functional pathways, thus providing a way to understand functions of genes for which information has not been previously available.

Finally, relationships among genes or annotations and sample conditions can be detected also by exploiting the association rule mining techniques, which extract correlations among dataset attributes. This technique is also used to analyze time-series microarray data to discover gene regulatory networks.

In this review, application of data mining techniques on microarray data is focused, with the aim of making researchers aware of the benefits of such techniques when analyzing microarray data. The chapter is organized as follows. The first two sections provide a description of microarray data, to highlight the issues concerned with their analysis, and a brief discussion about the data cleaning approaches that can be exploited to prepare data before data mining. The following four sections provide a survey of classification, feature selection, clustering and association rule mining techniques based on their aims and characteristics. Finally, the last two sections describe new trends and provide some prospects of data mining application to microarray data.

DATA MINING TECHNIQUES FOR MICROARRAY

Microarray Datasets

A microarray dataset E can be represented in the form of a gene expression matrix, in which each row represents a gene and each column represents a sample. For each sample, the expression level of all the genes under consideration is measured. Element e_{ij} in E is the measurement of the expression level of gene i for sample j, where $i=1,...,N$, $j= 1, ..., M$ and usually $N >> M$. Each sample is also characterized by a class label, representing the clinical situation of the patient or the biological condition of the tissue. The domain of class labels is characterized by C different values and label l_j of sample j takes a single value in this domain.

The format of a microarray dataset conforms to the normal data format of machine learning and data mining, where a gene can be regarded as a feature or attribute and a sample as an instance or a data point. However, the main characteristics of this data type are the high number of genes (usually tens of thousands) and the low number of samples (less then one hundred). This peculiarity causes specific challenges in analyzing microarray data (e.g., complex data interactions, high level of noisy, lack of biological absolute knowledge)