Chapter 60

Information Extraction from Microarray Data:
A Survey of Data Mining Techniques

Alessandro Fiori
Institute for Cancer Research and Tretment, Italy

Francesco Gavino Brundu
Politecnico di Torino, Italy

Alberto Grand
Institute for Cancer Research and Tretment, Italy

Domenico Schioppa
University of Torino Medical School, Italy

Giulia Bruno
Politecnico di Torino, Italy

Andrea Bertotti
University of Torino Medical School, Italy

ABSTRACT

Nowadays, a huge amount of high throughput molecular data are available for analysis and provide novel and useful insights into complex biological systems, through the acquisition of a high-resolution picture of their molecular status in defined experimental conditions. In this context, microarrays are a powerful tool to analyze thousands of gene expression values with a single experiment. A number of approaches have been developed to detecting genes highly correlated to diseases, selecting genes that exhibit a similar behavior under specific conditions, building models to predict disease outcome based on genetic profiles, and inferring regulatory networks. This paper discusses popular and recent data mining techniques (i.e., Feature Selection, Clustering, Classification, and Association Rule Mining) applied to microarray data. The main characteristics of microarray data and preprocessing procedures are presented to understand the critical issues introduced by gene expression values analysis. Each technique is analyzed, and relevant examples of pertinent literature are reported. Moreover, real use cases exploiting analytic pipelines that use these methods are also introduced. Finally, future directions of data mining research on microarray data are envisioned.

DOI: 10.4018/978-1-4666-9562-7.ch060

INTRODUCTION

The last few years have witnessed the explosive growth of biological data, with the development of new technologies and revolutionary changes in biomedicine and biotechnologies. The analysis of gene expressions retrieved with DNA microarray technology has become a fundamental tool in genomic research. Since microarray data show high levels of noise, high dimensionality and small sample data sets, data cleaning and data mining approaches have become fundamental to extracting relevant biological and genetic knowledge from this kind of data. Various data mining techniques can be applied, grouped into four vast categories: Feature Selection, Clustering, Classification and Association Rule Mining.

Since genetic data are noisy and can be affected by technical differences (e.g., hybridization parameters), it is usually necessary to employ data cleaning procedures before applying any data mining algorithm. In particular, normalization of expression values in a defined range and batch effect[1] removal are usually performed. Moreover, microarray data can carry some redundancy, as they include probes/genes that do not contain relevant information for the problem. Feature Selection techniques are dimensionality reduction methods applied prior to analysis to identify and remove redundant and useless features. The feature selection algorithms applied to microarray data allow identifying genes that are highly correlated with disease categories. In the same disease type, sets of genes usually show similar expression values. Clustering techniques attempt to identify these patterns and define groups of samples that show similar expression profiles. Differently, classification is a procedure used to predict group membership for data instances. Given a training set of samples with a number of attributes (or features) and a class label (e.g., a phenotype characteristic), a model is created for the classes. Next, the model is exploited to assign an appropriate class label to new data. Classification methods are very useful in verifying that the same expression profiles retrieved from training data can also be identified in other datasets provided by different studies. Finally, relationships among genes and sample annotations can also be detected by exploiting 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 principle, traditional data mining techniques could be directly applied to microarray data. In many cases, however, researchers have adapted these methods to handle microarray characteristics and extract relevant knowledge from a biological point of view. Indeed, since microarrays are used to analyze the expression values of thousand of genes with a single experiment, several subsets of genes have similar behavior and are correlated under the same conditions (e.g., experimental conditions, disease). Thus, mining approaches proposed in the literature deal with these aspects to improve the quality of analysis results. Moreover, datasets built using different experimental data can suffer from a high level of noise, missing values and batch effects, due to technical failures and the heterogeneity of experimental procedures. For these reasons, ad-hoc preprocessing techniques have been proposed to solve these problems, by reducing the distortion in the original data. Finally, the imbalance between the high number of genes (usually tens of thousands) and the low number of samples (less than one hundred) introduces new challenges in terms of computational costs and result accuracy.

In this survey, we present data mining techniques developed for the analysis of microarray data with the aim of making researchers aware of the benefits and disadvantages of such techniques. Moreover, real use cases are presented to show how different approaches can be combined to extract relevant biological and medical knowledge. The chapter is organized as follows. The "Data preprocessing" section provides a description of