Category: Bioinformatics

Knowledge Discovery from Genomics Microarrays

Lei Yu
Binghamton University, USA

INTRODUCTION

The advent of genomic microarray technology enables simultaneously measuring the expressions of thousands of genes in massive experiments, and hence provides scientists, for the first time, the opportunity of observing complex relationships between various genes in a genome. In order to extract biologically meaningful insights from a plethora of data generated from microarray experiments, knowledge discovery techniques, which discover patterns, statistical or predictive models, and relationships among massive data, have been widely applied in microarray data analysis. For example, clustering can be applied to identify groups of genes that are regulated in a similar manner under a number of experimental conditions or groups of samples that show similar expression patterns across a number of genes (Jiang, Tang, & Zhang, 2004). Classification can be performed to characterize the cellular difference between different samples, such as between normal and cancer cells or between cancer cells with different responses to treatment, and can potentially be used to predict the classes of samples based on their gene expression patterns (Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005). Feature selection or gene selection can help identify among thousands of genes a small fraction of genes that are relevant for discriminating between different sample types, and may potentially lead to the identification of a few biologically relevant "marker" genes for subsequent biological validation (Saeys, Inza, & Larranaga, 2007).

This article provides a brief introduction to the field of knowledge discovery and its applications in discovering useful knowledge from genomic microarray data. It describes common knowledge discovery tasks for genomic microarray data, presents representative methods for each task, and identifies emerging challenges and trends in knowledge discovery from genomic microarray data.

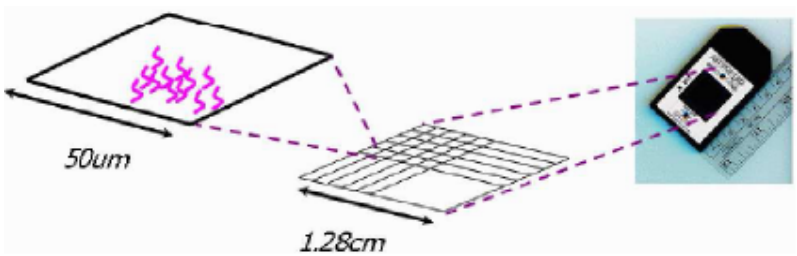
BACKGROUND

Knowledge discovery from data refers to the overall process of converting raw data into useful information, which consists of data preprocessing, data mining, and postprocessing of data mining results (Tan, Steinbach, & Kumar, 2005). The purpose of data preprocessing is to transform the raw input data into an appropriate format for subsequent data mining process. The tasks involved in data preprocessing include cleaning data to remove noise, duplicate, inconsistent, or missing information, integrating data from multiple sources, transforming data values into the right scale and format, and selecting instances and features that are relevant to the data mining task at hand. Data mining is the automatic process of extracting interesting patterns or knowledge from data. Data mining tasks can be generally divided into two major categories: predictive tasks and descriptive tasks. The objective of predictive tasks is to predict the values of a particular feature, based on the values of other features. The objective of descriptive tasks is to derive patterns that summarize the underlying relationships in data. These tasks are often exploratory in nature and frequently require postprocessing techniques which validate and explain the results. For example, visualizing the patterns allows analysts to explore the result from multiple viewpoints. Statistical tests can be used to validate the significance of the results and eliminate patterns that are generated by chance. For a comprehensive discussion on various knowledge discovery tasks, please refer to widely adopted text books on data mining (Han & Kamber, 2005; Tan et al., 2005). This article focuses on knowledge discovery tasks that are commonly performed on genomic microarray data introduced next.

Gene expression microarrays are silicon chips that simultaneously measure the mRNA expression levels of thousands of genes. Different types of microarrays use different technologies for constructing these chips and measuring gene expression levels. Detailed description of these technologies is beyond the scope of this article. Interested readers can refer to Draghici's book (Draghici, 2003) for an introduction. Here we briefly describe microarray technologies using Affymetrix arrays (shown in Figure 1) as an example, which are currently one of the most popular commercial arrays. However, the methodology for constructing other types of arrays would be similar, but would use different technology-specific data preparation and cleaning steps (Piatetsky-Shapiro & Tamayo, 2003).

Each Affymetrix array (GeneChip) contains probes for different genes tiled in a grid-like fashion. The simultaneous measure of expression levels of thousands of genes is done by hybridizing a complex mixture of mRNAs (derived from tissue or cells) to the probes. Hybridization events

Figure 1. An example of microarray: Affymetrix GeneChip (right), its grid (center), and a cell in the grip (left). Image courtesy of Affymetrix



are detected using a fluorescent dye and a scanner that can detect fluorescence intensities. The scanner and associated software perform various forms of image analysis to measure and report raw gene expression values. This allows for a quantitative readout of gene expression on a gene-by-gene basis (Piatetsky-Shapiro & Tamayo, 2003). To date, one microarray chip is capable of measuring expression levels for over 40,000 genes in the entire human genome.

The expression level of a specific gene among thousands of genes measured in an experiment is eventually recorded as a numerical value. Expression levels of the same set of genes under study are normally accumulated through multiple experiments on different samples (or the same sample under different conditions) and recorded in a data matrix. In data mining, data is often stored in the form of a matrix, of which each column is described by a feature or attribute and, each row consists of feature-values and forms an instance, also named as a record or data point, in a multidimensional space defined by the features. Figure 2 illustrates two ways of representing microarray data in a matrix form. In Figure 2 (a), each feature is a sample (*S)* and each instance is a gene (*G*). For each gene, its expression levels are measured across all the samples (or conditions), so f_{ij} is the measurement of the expression level of the i^{th} gene for the j^{th} sample where $i = 1,..., n$ and $j = 1,..., m$. In Figure 2 (b), the data matrix is the transpose of the one in Figure 2 (a), in which features are genes and instances are samples. Sometimes, data in Figure 2 (b) may have class labels c_i for each instance, represented in the last column. The class labels can be different cellular conditions of the underlying samples. A typical microarray data set may contain thousands of genes but only a small number of samples (often less than a hundred). The number of samples is likely to remain small at least for the near future due to the expense of collecting microarray samples.

Different data mining tasks can be performed on the two different forms of data shown in Figure 2. When genes are treated as instances (as in Figure 2 (a)), *gene clustering* can be performed to find similarly expressed genes across many samples. When samples are treated as instances (as in Figure 2 (b)), three different tasks can be performed: *sample clustering* which aims to group similar samples together and discover classes or subclasses of samples, *sample classification* which aims to classify diseases or phenotypes of novel samples based on patterns learned from training

Figure 2. Two views of a microarry data matrix

	S_1	S_2	. . .	S_m
G_1	f_{11}	f_{12}	. . .	f_{1m}
G_2	f_{21}	f_{22}	. . .	f_{2m}
.	.	.	.	.
.	.	.		.
.	.	.	.	.
.	.	.		.
.	.	.	.	.
G_n	f_{n1}	f_{n2}	. . .	f_{nm}

	G_1	G_2	. . . . .	G_n	
S_1	f_{11}	f_{21}	. . . . .	f_{n1}	c_1
S_2	f_{12}	f_{22}	. . . . .	f_{n2}	c_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
S_m	f_{1m}	f_{2m}	. . . . .	f_{nm}	c_m