Heterogeneous Gene Data for Classifying Tumors

Benny Yiu-ming Fung
The Hong Kong Polytechnic University, Hong Kong

Vincent To-yee Ng
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

When classifying tumors using gene expression data, mining tasks commonly make use of only a single data set. However, classification models based on patterns extracted from a single data set are often not indicative of an entire population and heterogeneous samples subsequently applied to these models may not fit, leading to performance degradation. In short, it is not possible to guarantee that mining results based on a single gene expression data set will be reliable or robust (Miller et al., 2002). This problem can be addressed using classification algorithms capable of handling multiple, heterogeneous gene expression data sets. Apart from improving mining performance, the use of such algorithms would make mining results less sensitive to the variations of different microarray platforms and to experimental conditions embedded in heterogeneous gene expression data sets.

BACKGROUND

Recent research into the mining of gene expression data has operated upon multiple, heterogeneous gene expression data sets. This research has taken two broad approaches, addressing issues related either to the theoretical flexibility that is required to integrate gene expression data sets with various microarray platforms and technologies (Lee et al., 2003), or – the focus of this chapter - issues related to tumor classification using an integration of multiple, heterogeneous gene expression data sets (Bloom et al., 2004; Ng, Tan, & Sundarajan, 2003). This type of tumor classification is made more difficult by three types of variation, variation in the available microarray technologies, experimental and biological variations, and variation in the types of cancers themselves.

The first type of variation is caused by different probe array notations of available microarray technologies. The two most common microarray technologies are photolithographically synthesized oligonucleotide probe arrays and spotted cDNA probe arrays. These have both been reviewed by Sebastiani, Gussoni, Kohane, and Ramoni (2003). They differ in their criteria for measuring gene expression levels. Oligonucleotide probe arrays measure mRNA abundance indirectly, while spotted cDNA probe arrays measure cDNA relative to hybridized reference mRNA samples. With the two common microarray technologies, there exist different probe array notations (Lee et al., 2003). For example, human probe array notations include GeneChip® (Affymetrix, *Santa Clara, CA*) U133, U95, and U35 accession number sets, BMR chips (Stanford University), UniGene clusters, cDNA clone ID and GenBank identifiers. Although the notations used in different technologies sometimes referred to the same set of genes, this does not indicate a simple one-to-one mapping (Ramaswamy, Ross, Lander, & Golub, 2003). Users of these notations should be aware of the potential for duplicated accession numbers in mapped results.

Another type of variation, the statistical variation among different gene expression data sets is unavoidable because experimental and biological variations are embedded in data sets (Miller et al., 2002). First of all, individual gene expression data sets are conducted by different laboratories with different experimental objectives and conditions even when using the same microarray technology. Integration of them is a painful task. Secondly, the expression levels of genes in experiments are normally measured by the ratio of the expression levels of the genes in the varying conditions of interest to the expression levels of the genes in some reference conditions. These reference conditions are varied from experiment to experiment. This is not a problem if sample sizes are large enough. Zien, Fluck, Zimmer, and Lengauer (2003) proposed that the use of larger sample sizes (e.g. 20 samples) can prevent mining results of gene expression data from suffering technical and biological variations, and produce more reliable results. Most gene expression data sets, however, contain fewer than 20 samples per class. A more flexible

solution would be to meta-analyze multiple, heterogeneous gene expression data sets, forming meta-decisions from a number of individual decisions.

The last difficulty is to find common features in various cancer types. These features can be referred as some sets of significant genes which are most likely expressed in most cancer types, but they may be expressed differently in varying cancer types. The study of human cancer has recently discovered that the development of antigen-specific cancer vaccines leads to the discovery of immunogenic genes. This group of tumor antigens has been introduced as the term "cancer-testis (CT) antigen" (Coulie et al., 2002). Discovered CT antigens are recently grouped into distinct subsets and named as "cancer/testis (CT) immunogenic gene families". Some works show that most CT immunogenic gene families are expressed in more than one cancer type, but with various expression frequencies. Currently, researchers have reviewed and summarized that the current discovery is 44 CT immunogenic gene families consisting of 89 individual genes in total (Scanlan, Simpson, & Old, 2004).

MAIN THRUST

It is possible is to make classification algorithms more reliable and robust by combining multiple, heterogeneous gene expression data sets. A simple combination method is to merge or append one data set to another. Unfortunately, this method is inflexible because data sets have various scales and ranges of variations. These are required to be the same in order to have consistent scales for comparisons after the combination.

In this chapter, we discuss two approaches to combine data sets consisting of variation in the available microarray technologies. The first, and simplest, approach is to normalize gene expression levels of genes in the data sets with mean zero and standard deviation one (i.e. standard normal distribution, *N(0, 1)*) according to the means and standard deviations across samples in individual data sets. While this approach is simple to apply, it assumes that all genes have the same or similar expression rates. However, this assumption is incorrect. The fact is that only a small subset of genes reflects the existence of tumors, and that the remaining genes in a tumor are not epidemiologically significant. It should also be noted that the reflected genes do not all express at the same rate. Therefore, when all genes in data sets are normalized to have *N(0, 1)*, the variations of the reflected genes may be underestimated and the variations of genes which are stable and irrelevant may be overestimated. This situation worsens as the number of genes in data sets increases.

The second and a better approach is to select a subset of reference genes, also known as significant genes, and to use the expression levels of these genes to estimate scaling factors which are used to rescale the expression levels of genes in other data sets with the same set of reference genes as in the original subset. This approach has two advantages. The first is that it allows the effects of outliers caused by non-significant genes to be eliminated while using only a subset of significant genes. In a gene expression data set, only a proportion of genes is tumor-specific. Because gene expression data contains high-dimensional data, by focusing on such tumor-specific genes in classification would reduce computational costs. The second advantage is that it improves the quality of the normalization or re-scaling since it avoids the underestimation of expression level of significant genes, a problem which may arise because of the presence of large amounts of non-significant genes. We also note that the selection algorithms are the focus of much current research. Some works that utilize existing features selection algorithms include Dudoit, Yang, Callow, and Speed (2002), Bloom et al. (2004), and Lee et al. (2003). New or enhanced algorithms have been proposed by Park et al. (2003), Ng, Tan, and Sundarajan. (2003), Choi, Yu, Kim, and Yoo (2003), Storey & Tibshirani (2003), Chilingaryan, Gevorgyan, Vardanyan, Jones, & Szabo (2002), and Golub et al. (1999).

In recent years, detection of significant genes was mainly done using fold-change detection. This detection method is unreliable because it does not take into account statistical variability. Currently, however, most algorithms that are used to select significant genes apply statistical methods. In the rest of the chapter, we first present some recent works on the identification of significant genes using statistical methods. We then briefly describe our proposed measure, Impact Factors (IFs), which can be used to carry out tumor classification using heterogeneous gene expression data (Fung & Ng, 2003).

Statistical Methods

The most common statistical method for identifying significant genes is the two-sample *t*-test (Cui & Churchill, 2003). The advantage of this test is that, because it requires only one gene to be studied for each *t*-test, it is insensitive to heterogeneity in variance across a couple of genes. However, while reliable *t*-values require large sample sizes, gene expression data sets normally consist of small sample sizes. This problem of small sample sizes can be overcome using global *t*-tests, but it assumes that the variance is homogeneous between different genes (Tusher, Tibshirani, & Chu, 2001). Tusher, Tibshirani, and Chu (2001) proposed a