Expert Systems with Applications 36 (2009) 9072–9081

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An expert system to classify microarray gene expression data using gene selection by decision tree

Jorng-Tzong Horng [a,b,c,*], Li-Cheng Wu [b], Baw-Juine Liu [d], Jun-Li Kuo [a], Wen-Horng Kuo [e], Jin-Jian Zhang [e]

[a] Department of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan
[b] Institute of System Biology and Bioinformatics, National Central University, Taiwan
[c] Department of Bioinformatics, Asia University
[d] Department of Computer Science and Information Engineering, Yuan Ze University, Taiwan
[e] College of Medicine, National Taiwan University, Taiwan

ARTICLE INFO

Keywords:
Expert system
Machine learning
Bioinformatics
Microarray gene expression
Decision tree

ABSTRACT

Gene selection can help the analysis of microarray gene expression data. However, it is very difficult to obtain a satisfactory classification result by machine learning techniques because of both the curse-of-dimensionality problem and the over-fitting problem. That is, the dimensions of the features are too large but the samples are too few. In this study, we designed an approach that attempts to avoid these two problems and then used it to select a small set of significant biomarker genes for diagnosis. Finally, we attempted to use these markers for the classification of cancer. This approach was tested the approach on a number of microarray datasets in order to demonstrate that it performs well and is both useful and reliable.

1. Introduction

According to a report from Department of Health, Executive Yuan, ROC (Taiwan), cancer was the leading cause of death in the Taiwan in 2002. It is also the fourth most common disease and the second leading cause of death in USA. Up to the present, no effective medical treatments are available for most known cancers. The aim when treating a malignant tumor is to kill the tumor in order to stop the spread of the cancer. Therefore may different approaches have been tried to help understand cancer (Su et al., 2001). One useful method is microarray expression studies combined with computer based analysis. This approach allows scientists to discover and understand more about various important features of cancer (Antonov et al., 2004).

Microarray studies are a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern and it is used to discover which specific genes are important to the development of a disease (www.ncbi.nlm.nih.gov/About/primer/microarrays.html). The expression level of the thousands of genes in a cell can be measured simultaneously. Therefore microarray studies enable clinicians and biologists to obtain the gene expression profile of a given tissue sample rapidly and compare it with other samples (Brown, 2002; Wang et al., 2005). A topic of great interest is the analysis of gene expression data associated with a specific diagnosis. For example, the study of expression profiles between microarray samples from cancer patients and normal subjects, allowing these genes to be classified based on differences in expression levels (Qiu, Wang, & Liu, 2005).

Most studies consider that the functioning of cells is understandable by observing gene expression (Aronow, Richardson, & Handwerger, 2001), and the topic of cancer diagnosis combined with microarray experiments has been discussed extensively (Antonov et al., 2004; Choi et al., 2005). As a result, analyzing microarray gene expression data has become a novel approach to cancer prognosis (Brennan et al., 2005).

However, there are some major technical difficulties or problems that confront researchers in this area. For example, genetic variability affects gene expression. That is, the expression levels of two patients with the same disease may differ significantly (Li, Zhang, & Ogihara, 2004). Additionally, there are many noise factors that affect microarray gene expression datasets and how to filter out noise is a thorny problem that must be solved (Li et al., 2004).

Computational analysis and computing can help researchers to collate a group of signature genes for a certain disease (Bae & Mallick, 2004; Buturovic, 2006; Wang et al., 2005; Yeung, Bumgarner, & Raftery, 2005). Owing to the high price of microarray chips and a lack of tissues from patients, datasets are too few in number to use machine learning. In addition, the processing and material used for microarray analysis differ between manufacturers and so it is difficult to identify a unique set of genes that can form an integrated dataset (Ein-Dor et al., 2005). Therefore, while the number of samples for each category is usually balance for the computer analysis, the ratio of cancer patients to normal adults is much smaller in the real world.

* Corresponding author. Address: Department of Computer Science, National Central University, Taiwan.
E-mail address: horng@db.csie.ncu.edu.tw (J.-T. Horng).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.12.037

Table 1
Comparison of some public systems or approaches to analyzing microarray data.

System name	System characteristics			Classification methods				Publication
	Repeating test	Gene clustering	Gene selection	C4.5	SVM	K-NN	NB	
			√	√	√	√		Li, J. *Bioinformatics*, 2003
HykGene	√	√	√	√	√	√	√	Wang, Y. *Bioinformatics*, 2004
GEMS	√		√		√	√		Statnikov, A. *Bioinformatics*, 2005
BMA			√		√	√		Yeung, K. Y. *Bioinformatics*, 2005
	√		√		√	√	√	Antonov, A. V. *Computational Biology and Chemistry*, 2005
PCP			√		√			Buturovic, L. J. *Bioinformatics*, 2006

One challenge is classifying or predicting diagnostic categories using microarray data because the number of genes is always much greater than the number of tissue samples (Yeung et al., 2005). How to choose a small and discriminative subset of genes from among tens of thousands of genes is very difficult. Therefore, gene selection becomes the most necessary prerequisite for a diagnostic classifying system. However, the best combination of classification and gene selection is understood poorly, because there is another methodological trouble associated with training microarray data. This is the problem of "over-fitting" (Statnikov et al., 2005). Briefly speaking, over-fitting means that one can obtain good performance using a training set, but when new data is used, a satisfactory result cannot be obtained using the trained model. This occurs often when there are a small number of high-dimension samples.

The common procedures used to classify microarray data have been described previously (Antonov et al., 2005; Buturovic, 2006; Chu et al., 2005; Qiu et al., 2005; Statnikov et al., 2005; Wang et al., 2005) and are as follows:

1. Raw microarray data normalization. The tens of thousands of probes on the microarray provide weak or strong signals after an experiment. Normalization converts the signals to variables, and appropriate normalization is used to filter out some noise.
2. Gene selection. In order to solve the over-fitting problem, gene selection becomes critical. Gene selection includes gene filtering, gene clustering, gene ranking and gene extraction. Some basic numerical or statistical analysis, such as *t*-test (Fisher, 1932), *F*-score (Golub et al., 1999) or standard deviation (SD) Seigel, 2003 are applied to filtering the genes as a pre-procedure. The reason for gene selection has been pointed out in many previous studies, namely that it reduces the dimensions and improves performance during classification (Antonov et al., 2005; Buturovic, 2006; Li et al., 2003; Statnikov et al., 2005; Wang et al., 2005; Yeung et al., 2005).
3. Classification. It is essential to choose the most appropriate and suitable machine learning classification algorithm by testing all possible procedures using these systems because good gene selection ought to improve classification performance. These approaches include C4.5 decision tree (C4.5) (Quinlan, 1993),

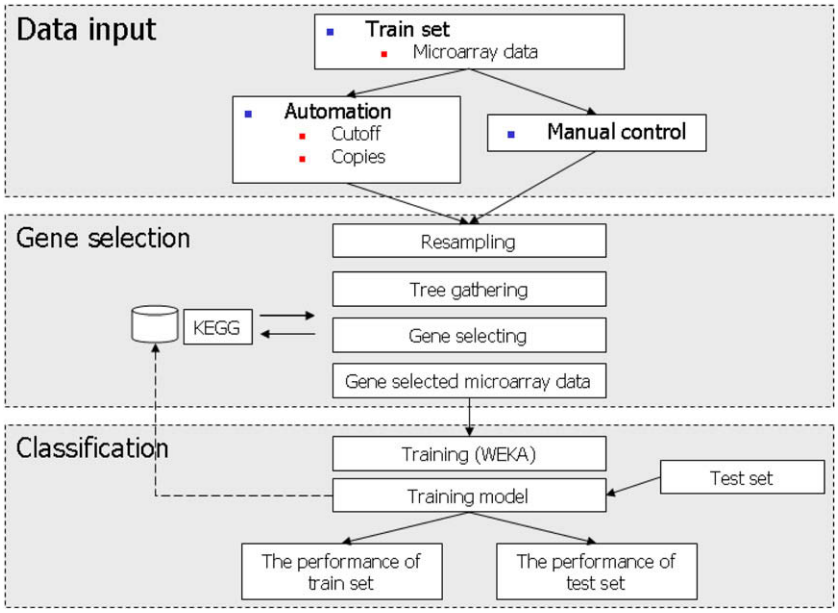

Fig. 1. System flow for selection of marker genes.