Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 178 (2008) 88–105
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Direct integration of microarrays for selecting informative genes and phenotype classification

Youngmi Yoon [a,c], Jongchan Lee [a], Sanghyun Park [a,*], Sangjay Bien [a], Hyun Cheol Chung [b], Sun Young Rha [b]

[a] Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea
[b] Department of Internal Medicine, Cancer Metastasis Research Center, Yonsei University College of Medicine, South Korea
[c] Department of Information Technology, Gachon University of Medicine and Science, South Korea

Received 7 February 2007; received in revised form 25 July 2007; accepted 1 August 2007

Abstract

The ability to provide thousands of gene expression values simultaneously makes microarray data very useful for phenotype classification. A major constraint in phenotype classification is that the number of genes greatly exceeds the number of samples. We overcame this constraint in two ways; we increased the number of samples by integrating independently generated microarrays that had been designed with the same biological objectives, and reduced the number of genes involved in the classification by selecting a small set of informative genes. We were able to maximally use the abundant microarray data that is being stockpiled by thousands of different research groups while improving classification accuracy. Our goal is to implement a feature (gene) selection method that can be applicable to integrated microarrays as well as to build a highly accurate classifier that permits straightforward biological interpretation. In this paper, we propose a two-stage approach. Firstly, we performed a direct integration of individual microarrays by transforming an expression value into a rank value within a sample and identified informative genes by calculating the number of swaps to reach a perfectly split sequence. Secondly, we built a classifier which is a parameter-free ensemble method using only the pre-selected informative genes. By using our classifier that was derived from large, integrated microarray sample datasets, we achieved high accuracy, sensitivity, and specificity in the classification of an independent test dataset.

Keywords: Data mining; Microarray data analysis; Microarray data integration; Microarray data classification; Informative gene selection

1. Introduction

Recently researchers have examined tumor cell specific gene expression patterns and have made use of the molecular characteristics of tumor tissue for diagnostic purposes. Since microarray technology is capable of

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park).

0020-0255/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2007.08.013