International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **1**, January **2010**

 ISSN 1349-4198
pp. 117–125

A THREE-STAGE METHOD TO SELECT INFORMATIVE GENES FOR CANCER CLASSIFICATION

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Michifumi Yoshioka[1] and Safaai Deris[2]

[1]Department of Computer Science and Intelligent Systems
Graduate School of Engineering
Osaka Prefecture University
Sakai, Osaka 599-8531, Japan
mohd.saberi@sig.cs.osakafu-u.ac.jp; { sigeru; yoshioka }@cs.osakafu-u.ac.jp

[2]Department of Software Engineering
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
81310 Skudai, Johore, Malaysia
safaai@utm.my

Received February 2009; revised August 2009

Abstract. *Microarray technology has provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment. One of the urgent issues in the use of microarray data is the selection of a small subset of genes from the thousands of genes in the data that contributes to a disease. This selection process is difficult due to many irrelevant genes, noisy genes, and the availability of the small number of samples compared to the huge number of genes (high-dimensional data). In this study, we propose a three-stage gene selection method to select a small subset of informative genes that is most relevant for the cancer classification. It has three stages: 1) pre-selecting genes using a filter method to produce a subset of genes; 2) optimising the gene subset using a multi-objective hybrid method to yield near-optimal gene subsets; 3) analysing the frequency of appearance of each gene in the different near-optimal gene subsets to produce a small subset of informative genes. The experimental results show that our proposed method is capable in selecting the small subset to obtain better classification accuracies than other related previous works as well as five methods experimented in this work. Additionally, a list of informative genes in the final gene subsets is also presented for biological usage.*

Keywords: A three-stage method, Gene selection, Hybrid approach, Microarray data

1. **Introduction.** The recent development of microarray technologies has enabled biologists to quantify the expression levels of thousands of genes in a single experiment. It finally produces microarray data. A comparison between the gene expression levels of cancerous and normal tissues can also be done. This comparison is useful to select those genes that might anticipate the clinical behaviour of cancers. Thus, there is a need to select informative genes that contribute to a cancerous state. An informative gene is useful for cancer classification. However, the gene selection process poses a major challenge because of the following characteristics of microarray data: the huge number of genes compared to the small number of samples (high-dimensional data), irrelevant genes and noisy data.

To overcome the challenge, a gene selection method is normally used to select a subset of genes that increases the classifier's ability to classify samples more accurately. Efficient