International Journal of Innovative
Computing, Information and Control
Volume **5**, Number **10(A)**, October **2009**

 ISSN 1349-4198
pp. **2959–2968**

A TWO-STAGE METHOD TO SELECT A SMALLER SUBSET OF 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 July 2008; revised December 2008

Abstract. *Gene expression data measured by microarray machines are useful for cancer classification. However, it faces with several problems in selecting genes for the classification due to many irrelevant genes, noisy data, and the availability of a small number of samples compared to a huge number of genes (high-dimensional data). Hence, this paper proposes a two-stage gene selection method to select a smaller (near-optimal) subset of informative genes that is most relevant for the cancer classification. It has two 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 automatically yield a smaller subset of informative genes. Three gene expression data sets are used to test the effectiveness of the proposed method. Experimental results show that the performance of the proposed method is superior to other experimental methods and related previous works.*

Keywords: Cancer classification, Filter method, Gene selection, Genetic algorithm, Gene expression data, Hybrid method

1. **Introduction.** Microarray technology is used to measure the expression levels of thousands of genes simultaneously, and finally produce gene expression 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. However, the gene selection poses a major challenge because of the following characteristics of gene expression data:

1) High-dimensional data, for example, a huge number of genes and a small number of samples are in the ranges of 7,000-15,000 and 30-200, respectively.
2) Most genes are not relevant for classifying different tissue types.
3) These data have noisy genes.

To overcome the problems, a gene selection method is used to select a subset of genes that maximises the classifier's ability to classify samples more accurately. The gene selection method has several advantages such as improving classification accuracy, reducing the dimensionality of data, and removing irrelevant and noisy genes.