International Journal of Innovative Computing, Information and Control
Volume **5**, Number **12(A)**, December **2009**

ICIC International ©2009 ISSN 1349-4198
pp. **4627–4635**

FEATURE SELECTION BY WEIGHTED-SNR FOR CANCER MICROARRAY DATA CLASSIFICATION

Supoj Hengpraprohm[1] and Prabhas Chongstitvatana[2]

[1]Faculty of Science and Technology
Nakhon Pathom Rajabhat University
Nakhon Pathom 73000, Thailand
supojn@yahoo.com

[2]Department of Computer Engineering
Chulalongkorn University
Bangkok 10330, Thailand
prabhas@chula.ac.th

Received July 2008; revised December 2008

Abstract. *Feature selection technique is widely used to improve the data analysis of high dimensional data especially in a classification task. Cancer microarray data classification task belongs to this category. Many researches studied the feature selection for microarray data classification. The major problem is that many feature selection methods must pre-define the number of feature. Unfortunately, the number of feature which is suitable is not known a priori. In this paper, we present a method to weight the value of each feature by SNR score. It is not necessary to pre-define the number of feature. Genetic Programming is employed as a classifier. The experimental results indicate that the proposed method yields good prediction accuracy.*
Keywords: Microarray data analysis, Cancer classification, Feature selection, Signal to noise ratio, Genetic programming

1. **Introduction.** The microarray technique is a popular method in bioinformatics. This technique allows us to study an organism in details. It can investigate thousands of genes simultaneously. The data of microarray consists of a small and high dimensional data. Therefore, it is very complex and difficult to analyze. The summary of the methods to microarray data analysis can be found in [1].

Cancer classification is a major challenging problem for microarray data analysis. The task is to identify the presence of cancer or to distinguish among specific cancers. Consequentially, a body of data has become established [2-7] and a number of classification tasks, by means of learning algorithms, are being tested for their accuracy on these data. Such researches aim to improve the effectiveness of the model derived from the learning algorithms [8-10]. The effectiveness of the model is measured by the classification accuracy on test data.

For large-scale dataset, any learning algorithm will consume a large computational resource. Also, performance and efficiency of the model may be decreased due to noise in data. There are many ways to alleviate these problems. For example, the number of sample can be reduced when the data is large [11]. In microarray data, dimensions of data should be reduced by feature selection. There are many researches that study feature selection methods [12-16]. Such methods aim to rank features by some scoring metric or finding a subset of features with respect to classifiers. However, the number of feature (gene) selected by scoring metrics must be pre-defined. In [16], we found that if the number of feature is unsuitable (too many or too few) the effectiveness of the learning