Computers in Biology and Medicine 41 (2011) 228–237

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# A hybrid feature selection method for DNA microarray data

Li-Yeh Chuang [a], Cheng-Huei Yang [b], Kuo-Chuan Wu [c], Cheng-Hong Yang [d,e,*]

[a] Department of Chemical Engineering, I-Shou University, Kaohsiung 80041, Taiwan
[b] Department of Electronic Communication Engineering, National Kaohsiung Marine University, Kaohsiung 81157, Taiwan
[c] Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 80708, Taiwan
[d] Department of Network Systems, Toko University, Chiayi 61363, Taiwan
[e] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 80708, Taiwan

ARTICLE INFO

*Article history:*
Received 21 August 2010
Accepted 8 February 2011

*Keywords:*
Feature selection
Taguchi-genetic algorithm
K-nearest neighbor
Leave-one-out cross-validation

ABSTRACT

Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. In cancer classification, available training data sets are generally of a fairly small sample size compared to the number of genes involved. Along with training data limitations, this constitutes a challenge to certain classification methods. Feature (gene) selection can be used to successfully extract those genes that directly influence classification accuracy and to eliminate genes which have no influence on it. This significantly improves calculation performance and classification accuracy. In this paper, correlation-based feature selection (CFS) and the Taguchi-genetic algorithm (TGA) method were combined into a hybrid method, and the K-nearest neighbor (KNN) with the leave-one-out cross-validation (LOOCV) method served as a classifier for eleven classification profiles to calculate the classification accuracy. Experimental results show that the proposed method reduced redundant features effectively and achieved superior classification accuracy. The classification accuracy obtained by the proposed method was higher in ten out of the eleven gene expression data set test problems when compared to other classification methods from the literature.



## 1. Introduction

Microarray data can provide valuable results for a variety of gene expression profile problems and contribute to advances in clinical medicine. The application of microarray data on cancer type classification has recently gained in popularity. Coupled with statistical techniques, gene expression patterns have been used to screen potential tumor markers. Differential expressions of genes are analyzed statistically and each gene expression is assigned to a certain category. The classification of gene expressions can substantially enhance the understanding of the underlying biological processes.

The goal of microarray data classification is to build an efficient and effective model that can differentiate the gene expressions of samples, i.e., determine normal or abnormal states, or classify tissue samples into different classes of diseases. The challenges posed in microarray classification are the limited amount of samples in comparison to the high-dimensionality of the sample, along with experimental variations in measured gene expression levels. In general, only a relatively small number of gene expression data show a strong correlation with a certain phenotype compared to the total number of genes investigated. This means that of the thousands of genes investigated only a small number show significant correlation with the phenotype in question. Thus, in order to analyze gene expression profiles correctly, feature (gene) selection is crucial for the classification process.

Recently, many gene expression data classification and gene selection techniques have been introduced. Kim et al. [1] proposed a novel method based on an evolutionary algorithm (EA) to assemble optimal classifiers and improve feature selection. Tang et al. [2] used an approach that selects multiple highly informative gene subsets. Wang et al. [3] proposed a new tumor classification approach based on an ensemble of probabilistic neural networks (PNN) and neighborhood rough set models based on gene selection. Shen et al. [4] proposed a modified particle swarm optimization that allows for the simultaneous selection of genes and samples. Xie et al. [5] developed a diagnosis model based on support vector machines (SVM) with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Li et al. [6] proposed an algorithm with a locally linear discriminant embedded in it to map the microarray data to a low dimensional space, while Huang et al. [7] proposed an improved decision forest for the classification of gene expression data that incorporates a built-in feature selection mechanism for fine-tuning.

* Corresponding author at: Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 807, Taiwan.
Tel.: + 886 7 381 4526x5639; fax: +886 7 383 6844.
*E-mail addresses:* chuang@isu.edu.tw (L.-Y. Chuang), kuo.chuan.wu@gmail.com (K.-C. Wu), chyang@cc.kuas.edu.tw (C.-H. Yang).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.02.004

In summary, the above feature selection methods can be divided into three common models: filter methods, wrapper methods, and embedded methods. The filter approach separates data before the actual classification process takes place and then calculates feature weight values, and thus features that accurately present the original data set can be identified. However, a filter approach does not account for interactions amongst the features. Methods in the filter approach category include correlation-based feature selection (CFS) [9], *t*-test, information gain [10], mutual information [11], and entropy-based methods [12]. Wrapper models, on the other hand, generally focus on improving classification accuracy of pattern classification problems and typically perform better (i.e., reach higher classification accuracy) than filter models. However, wrapper approaches are more computationally expensive than filter methods [13,14]. Several methods in this category have previously been used to perform feature selection of training and testing data, such as genetic algorithm (GA) [15], branch and bound algorithm [16], sequential search algorithm [17], tabu search [18,19], binary particle swarm optimization [20,21], and hybrid genetic algorithm [22]. Embedded techniques use an inductive algorithm. The inductive algorithm itself represents the feature selector and the classifier. Embedded techniques search for an optimal subset of features that is built into the classifier. Examples of these classification trees are ID3, C4.5 and random forest. The advantage of embedded algorithms is that they take the interaction with the classifier into account. A disadvantage of embedded algorithms is that they are generally based on a greedy mechanism, i.e., they only use top-ranked attributes to perform sample classification [8,23].

Many feature selection methods are combined with a local search process to improve accuracy. One example is presented in Oh et al. [22] who used a local search mechanism in their genetic algorithm. In this paper, we used the Taguchi method as a local search method embedded in the GA. The Taguchi method uses ideas from statistical experimental design to improve and optimize products, processes or equipment. The two main tools of the Taguchi method are: (a) the signal-to-noise ratio (SNR), which measures quality and (b) orthogonal arrays (OAs), which are used to simultaneously study the many design parameters involved. The Taguchi method is a robust design approach [24]. It has been successfully applied in machine learning and data mining, e.g., combined data mining and electrical discharge machining [20]. Sohn and Shin used the Taguchi experimental design for the Monte Carlo simulation of classifier combination methods [25]. Kwak and Choi used the Taguchi method to select features for classification problems [26]. Chen et al. optimized neural network parameters with the Taguchi method [27].

A hybrid feature selection approach consisting of two stages is presented in this study. In the first stage, a filter approach is used to calculate correlation-based feature weights for each feature, thus identifying relevant features. In the second stage, which constitutes a wrapper approach, the previously identified relevant feature subsets are tested by a Taguchi-genetic algorithm (TGA), which tries to determine optimal feature subsets. These optimal feature subsets are then appraised with the K-nearest neighbor method (KNN) [28,29] with leave-one-out cross-validation (LOOCV) [30,31] based on Euclidean distance calculations. Genetic algorithms [32,33] are utilized with randomness for a global search over the entire search space. The genetic operations crossover and mutation are performed to assist the search procedure in escaping from sub-optimal solutions [14]. In each iteration of the proposed nature-inspired method, the Taguchi method [24,34,35] is implemented to help explore better feature subsets (or solutions), which are somewhat different from those in the candidate feature subsets. In other words, the Taguchi algorithm is employed for a local search in the search space. Experimental results show that the proposed method achieved higher classification accuracy rates and outperformed the other methods from the literature it was compared to.

## 2. Material and methods

### 2.1. Correlation-based feature selection (CFS)

CFS was developed by Hall in 1999 [9]. CFS is a simple filter feature selection method that ranks feature subsets based on a correlation-based heuristic evaluation. This feature selection method is based on the following hypothesis:

> *Good feature subsets contain features highly correlated with* (*i.e., predictive of*) *the class, yet uncorrelated with* (*i.e., not predictive of*) *each other* [9].

This hypothesis is incorporated into the correlation-based heuristic evaluation equation as

$$Merit_S = \frac{k\overline{\gamma}_{cf}}{\sqrt{k+k(k-1)\overline{\gamma}_{ff}}} \quad (1)$$

where $Merit_S$ is the merit of a feature subset $S$ containing $k$ features, $\overline{\gamma}_{cf}$ is the average feature and class correlation, and $\overline{\gamma}_{ff}$ is the average feature-feature intercorrelation ($f \in S$).

General filter methods estimate the significance of a feature individually. CFS is then used to select the best combination of attribute subsets via score values from the original data sets. Heuristic search strategies are employed to identify the best combination. Common strategies include forward selection, backward elimination, and the best-first method. In this study, we used Weka [36] to implement CFS, and used the selected gene subsets to identify different cancer types and various diseases.

### 2.2. Genetic algorithm

A genetic algorithm (GA) was first developed by Holland in 1970. A GA is a stochastic search algorithm modeled on the process of natural selection underlying biological evolution. GAs has been successfully applied to many search, optimization, and machine learning problems [37]. They represent an intelligent exploitation of a random search within a defined search space to solve a problem. A GA proceeds in an iterative manner by generating new populations of strings from old ones. Every string is the encoded binary, real, etc., version of a candidate solution. An evaluation function connects a fitness measure to every string, indicating its fitness for the problem. Standard GAs apply genetic operators such as selection, crossover, and mutation on an initially random population in order to compute an entire generation of new strings. Further details of GA mechanisms can be found in Holland [37].

For a feature subset selection problem, a possible solution in the solution space is a specific feature subset that can be encoded as a string of $n$ binary digits (or bits). Each feature is represented by binary digits with values 1 or 0, which identify whether the feature is selected or not selected in the corresponding feature subset, respectively. This process is called solution (or chromosome) encoding. For instance, in the 0100100010 string of ten binary digits (i.e., a solution or a chromosome) the features 2, 5, and 9 are selected in the corresponding feature subset.

In the first step of a general GA some solutions are randomly selected from the solution space as the initial set *CS* of candidate solutions. The number of solutions in *CS* is denoted the population size. When two parent solutions $p_1$ and $p_2$ are selected from *CS*, the crossover operation is applied to generate a corresponding offspring $q$. In other words, each feature $i$ of offspring $q$ is the