Applied Soft Computing 24 (2014) 773–780

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data

Kun-Huang Chen[a], Kung-Jeng Wang[a,*], Kung-Min Wang[b], Melani-Adrian Angelia[a]

[a] Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, ROC
[b] Department of Surgery, Shin-Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan, ROC

ARTICLE INFO

*Article history:*
Received 6 April 2013
Received in revised form
21 December 2013
Accepted 13 August 2014
Available online 27 August 2014

*Keywords:*
Cancer classification
Gene expression
Particle swarm optimization
C4.5

ABSTRACT

*Background:* The application of microarray data for cancer classification is important. Researchers have tried to analyze gene expression data using various computational intelligence methods.
*Purpose:* We propose a novel method for gene selection utilizing particle swarm optimization combined with a decision tree as the classifier to select a small number of informative genes from the thousands of genes in the data that can contribute in identifying cancers.
*Conclusion:* Statistical analysis reveals that our proposed method outperforms other popular classifiers, i.e., support vector machine, self-organizing map, back propagation neural network, and C4.5 decision tree, by conducting experiments on 11 gene expression cancer datasets.

## 1. Introduction

The rapid development of microarray technology enables researchers to analyze thousands of genes simultaneously and obtain important information about the cell's function. This particular information can be used in cancer diagnosis and prognosis. However, given the characteristics of gene expression data (i.e., high dimension, high noise, and small sample size), the gene selection process remains challenging. A method for choosing the important subset of genes with high classification accuracy is needed to overcome this challenge. Such method would not only save computational costs, but will also enable doctors to identify a small subset of biologically relevant genes with certain cancers and target only a small number of genes in designing less expensive experiments [25]. Moreover, a highly accurate method can also assist in early diagnosis and drug discovery for cancer patients [2].

The method of gene selection generally falls into one of the following three categories: the filter, wrapper, and embedded approaches. The filter approach collects the intrinsic characteristics of genes in discriminating the targeted phenotype class and usually employs statistical methods, such as mutual information, statistical tests (-test, -test), and Wilcoxon's rank test, to directly select feature genes [47]. This approach is easily implemented, but ignores the complex interaction between genes. The "wrapper" approach [41] aims at selecting a subset of feature genes, typically with an induction algorithm to search for an initial gene subset which can then be used for further evaluating new feature gene subsets. The wrapper method is usually superior to the filter one since it involves inter-correlation of individual genes in a multivariate manner. The wrapper method can automatically determine the optimal number of feature genes for a particular classifier. The embedded method is similar to the wrapper method, while multiple algorithms can be combined in the embedded method to perform feature subset selection (Kahavi and John, 1997; [26]). In the embedded method, genetic algorithms (GAs) [58] are generally used as the search engine for feature subset, while other classification methods, such as KNN/GA (K nearest neighbors/genetic algorithms) [23], GA-SVM (genetic algorithms-support vector machine) [24], and so forth, are used to select feature subset. Estimation of distribution algorithm (EDA) [42] is a general framework of GA. Compared to traditional GA that employs crossover and mutation operators to create new population, EDA creates new populations by using a statistical approach to estimate the probability distribution of all promising individual solutions for the previous generation. EDA can also explicitly take into account specific interactions among the variables. When EDA is used to search for feature subsets, classification methods, such as Support vector machine (SVM) [6,36,18,59,13], which can deal with the high-dimension data in a limited sample space, can be used to select feature subsets.

[*] Corresponding author. Tel.: +886 2 2737 6769; fax: +886 2 2 737 6344.
*E-mail address:* kjwang@mail.ntust.edu.tw (K.-J. Wang).

http://dx.doi.org/10.1016/j.asoc.2014.08.032
1568-4946/© 2014 Elsevier B.V. All rights reserved.

Particle swarm optimization (PSO) is a popular meta-heuristic algorithm developed by [19]. PSO has been widely applied in many fields to solve various optimization problems, including gene selection [25,2,8,45,30]. In PSO, a swarm of particles with randomly initialized positions would move toward the optimal position along the search path that is iteratively updated based on the best particle position and velocity. The position of a particle can be used to represent a candidate solution for the problem. Among them, C4.5 is a decision tree-based classifier listed in the top 10 most influential data-mining algorithms in the research community [55]. Decision trees were a linear method as simple to understand and interpret.

This study proposes a method using the PSO algorithm to optimize the classification accuracy achieved using the C4.5 classifier (denoted as PSOC4.5). This study combines PSO for its excellent search capabilities and C4.5 for its knowledge interpretation advantage. This proposed hybrid technique combining PSO with C4.5 classifier has not been previously investigated by previous researchers. The performance of our proposed method is evaluated by testing the proposed method on 11 micro array datasets, which consist of 1 dataset from cancer patients of the National Health Insurance Research Database in Taiwan [34] and 10 from the Gene Expression Model Selector [15]. Moreover, we compare the performance of our proposed method with other well-known classifier algorithms, i.e., SVM, self-organizing map (SOM), back propagation neural network (BPNN), and C4.5. A statistical test is used to show that the proposed method outperforms other well-known classifiers in terms of classification accuracy.

The rest of the paper is organized as follows. In Section 2, we review the gene selection classification problem and related studies. Section 3 introduces the PSO algorithm and C4.5 classifier as the proposed approach. In Section 4, we present our experimental results and its comparison with those of other methods. Finally, we conclude the study in Section 5.

## 2. Overview of gene selection classification

### 2.1. Gene selection

DNA microarray is a technology that allows researchers to measure the expression levels of thousands of genes simultaneously in a single experiment. The method is usually used to compare the gene expression levels in tissues under different conditions, such as wild type versus mutant, or healthy versus diseased [12]. Ref. [33] propose a method for gene microarray classification that combines different feature reduction approaches for improving classification performance using a support vector machine (SVM) as our classifier. Their experiments were performed using several different datasets, and our results (expressed as both accuracy and area under the receiver operating characteristic (ROC) curve) show the goodness of the proposed approach with respect to the state of the art. Park [38] proposes a new approach for inferring combinatorial Boolean rules of gene sets for a better understanding of cancer transcriptome and cancer classification. To reduce the search space of the possible Boolean rules, we identify small groups of gene sets that synergistically contribute to the classification of samples into their corresponding phenotypic groups (such as normal and cancer).

In gene selection, we select the most informative genes, which are most predictive of its related class for classification. The gene selection process includes gene filtering, gene clustering, gene ranking, and gene extraction. Some basic numerical or statistical analysis, such as *t*-test, F-score, and standard deviation (Std.), are applied in filtering genes at the pre-procedure. Gene selection leads to reduced dimensions and improves classification performance.

Given the quantity and complexity of the gene expression data, an expert is unlikely to compute and compare the $n \times m$ gene expression matrix manually. Thus, machine learning and other artificial intelligence techniques have been widely used to classify or characterize gene expression data [5,6].

### 2.2. Related works

Several researchers have utilized the PSO algorithm to propose solutions for gene selection problems. Alba compared the use of PSO and genetic algorithm (GA), both augmented with SVM, as the classifier for high-dimensional microarray data. A modified PSO, namely, Geometric PSO, has been proposed for comparison with the GA on six public cancer datasets [2]. Li proposed a gene selection method by combining PSO with a GA and adopted SVM as the classifier. Their proposed approach was tested on three benchmark gene expression datasets: leukemia colon cancer, and breast cancer data. The aim of their hybrid method was to guide the proposed algorithm to prevent it from becoming trapped in the local optima [25]. Ref. [30] proposed an improved binary PSO combined with an SVM classifier to select a near-optimal subset of informative genes relevant to cancer classification. In the method by Mohamad et al., the existing rule for updating the particle position and velocity was modified.

Recently, Zhao proposed a novel hybrid framework (NHF) for gene selection and classification of high-dimensional microarray data, which combines information gain (IG), F-score, GA, PSO, and SVM. Three main steps comprise their proposed method. The performance of their proposed method was compared with those of the PSO-based, GA-based, ant colony optimization-based, and simulated annealing (SA)-based methods on five benchmark data sets: leukemia, lung carcinoma, and colon, breast, and brain cancers. The numerical results and statistical analysis showed that their proposed approach is capable of selecting a subset of predictive genes from a large noisy dataset and can capture the correlated structure in the data. Moreover, NHF performs significantly better than the other methods in terms of prediction accuracy with a smaller subset of features [57]. Ref. [51] propose an alternative to the existing methods for functionally annotating genes. The methodology involves building of classification models, validation and graphical representations of the results and reduction of the dimensions of the dataset. Ref. [8] used PSO + 1NN for feature selection. The proposed method was tested on eight benchmark datasets from UC Irvine Machine Learning Repository and then applied to an actual case of obstructive sleep apnea. The experimental results showed that the proposed method is significantly better than BPNN, logistic regression (LR), SVM, and C4.5. Further, [53] utilized a hybrid method combining GA and SVM as the classifier to identify the optimal subset of micro array datasets. The result they obtained from their proposed method was superior to those obtained by microPred and miPred. Heckerling [39] used ANN and genetic algorithms to develop models, based on sets of best predictor variables, for detecting urinary tract infection among women with urinary complaints Ref. [22] proposed a hybrid GA and PSO approach for designing a fuzzy based expert system. In their method, GA is used to find the rules and PSO is used to tune the membership function. They found that the proposed approach generated a compact fuzzy system with high classification accuracy for all six gene expression data sets when compared with other approaches.

Thus, PSO has been successfully applied to solve the gene selection problem. Therefore, we propose a method adopting a combination of PSO and C4.5 decision tree. Moreover, this work investigates the performance of this hybrid type of method, which has seldom been investigated.