Artificial Intelligence in Medicine (2006) **36**, 43–58

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming

Jin-Hyuk Hong, Sung-Bae Cho *

Department of Computer Science, Yonsei University, 134 Sinchon-dong, Sudaemoon-ku, Seoul 120-749, Republic of Korea

Received 21 December 2004; received in revised form 11 May 2005; accepted 17 June 2005

KEYWORDS
Genetic programming;
Ensemble;
Diversity;
Classification

Summary

Object: The classification of cancer based on gene expression data is one of the most important procedures in bioinformatics. In order to obtain highly accurate results, ensemble approaches have been applied when classifying DNA microarray data. Diversity is very important in these ensemble approaches, but it is difficult to apply conventional diversity measures when there are only a few training samples available. Key issues that need to be addressed under such circumstances are the development of a new ensemble approach that can enhance the successful classification of these datasets.
Materials and methods: An effective ensemble approach that does use diversity in genetic programming is proposed. This diversity is measured by comparing the structure of the classification rules instead of output-based diversity estimating.
Results: Experiments performed on common gene expression datasets (such as lymphoma cancer dataset, lung cancer dataset and ovarian cancer dataset) demonstrate the performance of the proposed method in relation to the conventional approaches.
Conclusion: Diversity measured by comparing the structure of the classification rules obtained by genetic programming is useful to improve the performance of the ensemble classifier.

1. Introduction

The classification of cancer is a major research area in the medical field. Such classification is an important step in determining treatment and prognosis [1,2]. Accurate diagnosis leads to better treatment and toxicity minimization for patients. Current morphological and clinical approaches that aim to classify tumors are not sufficient to recognize all the various types of tumors correctly. Patients may suffer from different type of tumors, even though they may show morphologically similar symptoms.

* Corresponding author. Tel.: +82 2 2123 2720; fax: +82 2 365 2579.
E-mail address: sbcho@cs.yonsei.ac.kr (S.-B. Cho).

0933-3657/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2005.06.002

A disease like a tumor is fundamentally a malfunction of genes, so utilizing the gene expression data might be the most direct diagnosis approach [1].

DNA microarray technology is a promising tool for cancer diagnosis. It generates large-scale gene expression profiles that include valuable information on organization as well as cancer [3]. Although microarray technology requires further development, it already allows for a more systematic approach to cancer classification using gene expression profiles [2,4].

It is difficult to interpret gene expression data directly. Thus, many machine-learning techniques have been applied to classify the data. These techniques include the artificial neural network [5—8], Bayesian approaches [9,10], support vector machines [11—13], decision trees [14,15], and *k* nearest neighbors [16].

Evolutionary techniques have also been used to analyze gene expression data. The genetic algorithm is mainly used to select useful features, while the genetic programming is used to find out a classification rule. Li et al. proposed a hybrid model of the genetic algorithm and *k* nearest neighbors to obtain effective gene selection [16], and Deutsch investigated evolutionary algorithms in order to find optimal gene sets [17]. Karzynsci et al. proposed a hybrid model of the genetic algorithm and a perceptron for the prediction of cancer [18]. Langdon and Buxton applied genetic programming for classifying DNA chip data [19]. Ensemble approaches have been also attempted to obtain highly accurate cancer classification by Valentini [20], Park and Cho [21], and Tan and Gilbert [22].

Highly accurate cancer classification is difficult to achieve. Since gene expression profiles consist of only a few samples that represent a large number of genes, many machine-learning techniques are apt to be over-fitted. Ensemble approaches offer increased accuracy and reliability when dealing with such problems. The approaches that combine multiple classifiers have received much attention in the past decade, and this is now a standard approach to improving classification performance in machine-learning [23,24]. The ensemble classifier aims to generate more accurate and reliable performance than an individual classifier. Two representative issues, which are ''how to generate diverse base classifiers'' and ''how to combine base classifiers'' have been actively investigated in the ensemble approach.

The first issue ''how to generate diverse base classifiers'' is very important in the ensemble approach. As already known, ensemble approaches that use a set of same classifiers offer no benefit in performance to individual ones. Improvement might be obtained only when the base classifiers are complementary. Ideally, as long as the error of each classifier is less than 0.5, the error rate might be reduced to zero by increasing the number of base classifiers. However, the results are different in practical experiments, since there is a trade-off between diversity and individual error [25]. Many researchers have tried to generate a set of accurate as well as diverse classifiers. Generating base classifiers for ensemble approaches is often called ensemble learning. There are two representative ensemble-learning methods: bagging and boosting [26].

Bagging (bootstrap aggregating) was introduced by Breimen. This method generates base classifiers by using a randomly organized set of samples from the original data. Bagging tries to take advantage of the randomness of machine-learning techniques. Boosting, introduced by Schapire, produces a series of base classifiers. A set of samples is chosen based on the results of previous classifiers in the series. Samples that were incorrectly classified by previous classifiers are given further chances to be selected to construct a training set. Arching and Ada-Boosting are currently used as promising boosting techniques [25,26].

Various other works have been used in an attempt to generate diverse base classifiers. Webb and Z heng proposed a multistrategy ensemble-learning method [25], while Optiz and Maclin provided an empirical study on popular ensemble methods [26]. Bryll et al. introduced attribute bagging, which generates diverse base classifiers using random feature subsets [24]. Islam et al. trained a set of neural networks to be negatively correlated with each other [27]. Other works have tried to estimate diversity and to select a subset of base classifiers for constructing an ensemble classifier [28,29].

The second issue ''how to combine base classifiers'' is important together with the first one. Once base classifiers are obtained, a choice of a proper fusion strategy can maximize the ensemble effect. There are many simple combination strategies, including majority vote, average, weighted average, minimum, median, maximum, product, and Borda count. These strategies consider only the current results of each classifier for a sample. Instead, other combination strategies (such as Naïve Bayes, behavior-knowledge space, decision templates, Dempster—Shafer combination, and fuzzy integral) require a training process to construct decision matrices. On the other hand, the oracle strategy, which requires only one classifier to classify a sample correctly, is often employed to provide a possible upper bound on improvement to classification accuracy.