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

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.

There has been much research on combination strategies. Verikas et al. comparatively tested various fusion methods on several datasets [30], and Kuncheva provided a formula for classification errors in simple combination strategies [23]. Tax compared averaging and multiplying as combining multiple classifiers [31], while Alexandre et al. compared sum and product rules [32]. Decision templates strategy, which was proposed by Kuncheva et al., has been compared with conventional methods [33]. Kim and Cho applied fuzzy integration of structure adaptive self-organizing maps (SOMs) for web content mining [34]. Shipp and Kuncheva tried to show relationships between combination methods and measures of diversity [29].

In this paper, we would like to address diversity in ensemble approaches and propose an effective ensemble approach by considering further diversity in genetic programming. A set of classification rules was generated by genetic programming, and then diverse ones were selected from among them in order to construct an ensemble classifier. In contrast to the conventional approaches, diversity was measured by matching the structure of the rules based on the interpretability of genetic programming. The paper also examines several representative feature selection methods and combination methods. Three popular gene expression datasets (lymphoma cancer dataset, lung cancer dataset and ovarian cancer dataset) were used for the experiments.

2. Ensemble genetic programming

Genetic programming was proposed by Koza in order to automatically generate a program that could solve a given problem [35]. It was originally similar to the genetic algorithm in many ways, but it was different in representation. An individual was represented as a tree composing of functions and terminal symbols. Various functions and terminal symbols were developed for the target application, and classification was one of the goals of genetic programming.

There are several works on ensemble approaches that use genetic programming. Zhang and Bhattacharyya used genetic programming for classifying the connection data of a simulated US Air Force LAN [36]. Brameier and Banzhaf evolved a set of classifiers by using genetic programming and combined them with several fusion methods [37], while Fernandez et al. studied the multi-population genetic programming empirically [38]. Imamura et al. proposed behavioral diversity in ensemble genetic programming [39].

Given training samples, functions, and terminal symbols, there is a set of decision boundaries that separate the data in the induced feature space F. This set of consistent hypotheses is called the version space [40]. Hypothesis f is in the version space if "$f(x_i) > 0$ and $y_i = 1$" or "$f(x_i) < 0$ and $y_i = -1$", where y_i indicates the class label of x_i. Classification using genetic programming can be regarded as searching for a hypothesis f that satisfies the condition with a given set of functions and terminal symbols.

Definition 1 (*A set of possible functions*). $\mathrm{Fu} = \{\mathrm{fu} \mid \mathrm{fu} \subset \{+, -, \times, \div\}\}$, where the functions can be extended according to applications.

Definition 2 (*A set of possible terminal symbols*). $\mathrm{TS} = \{\mathrm{ts} \mid \mathrm{ts} \subset \mathrm{Fe}\}$, where Fe is a given set of features $\{f_1, f_2, \ldots, f_n\}$.

Definition 3 (*A set of possible hypotheses*). $H = \{f \mid f(x_i) = t_{(\mathrm{fu},\ \mathrm{ts},\ \mathrm{depth})}(x)$, where $t \in T\}$, where our tree space T is simply equal to F.

Definition 4 (*The version space **V***). $V = \{f \in H \mid \forall i \in \{1, \ldots, m\} y_i f(x_i) > 0\}$, where m is the number of training samples.

Definition 5 (*The redefined version space **V′***). $V' = \{t \in T \mid y_i(t_{(\mathrm{fu},\ \mathrm{ts},\ \mathrm{depth})}(x_i)) > 0,\ i = 1, \ldots, m\}$.

Since H is a set of hyperplanes, there is a *bijection* between unit vectors t and hypotheses f in H. A version space only exists when the training samples are linearly separable in the feature space. Thus, linear separability of the training samples in the feature space is required. Since there is a duality between the feature space F and the tree space T (originally from the parameter space W in Vapnik's works), points in T correspond to hyperplanes in F. That is, if a training sample x_i is observed in the feature space, the set of separating hyperplanes is restricted to classify x_i correctly.

Definition 6 (*An ensemble hypothesis*). $\mathrm{EH} = \{\mathrm{eh} \mid \mathrm{eh}(x_i) = \mathrm{Majority\_vote}(t_1(x_i), \ldots, t_l(x_i))$, where $t_j \in T$, l is the ensemble size$\}$.

Definition 7 (*Volume of version space **V′***). $\mathrm{Vol}(t)$ = the size of version space V' that satisfies t.

Definition 8 (*Accuracy of **t***).

$$\mathrm{Acc}(x_i) = \frac{\mathrm{Vol}(y_i t_j(x_i) > 0)}{\mathrm{Vol}(y_i t_j(x_i) > 0) + \mathrm{Vol}(y_i t_j(x_i) < 0)}.$$

If a test sample x_i with label y_i is given, each t_j generates $t_{j(\mathrm{fu,\ ts,\ depth})}(x_i)$. When the majority vote is considered as the fusion strategy, eh may produce a correct result if there are more than $l/2$ t_j's that satisfy $y_i t_{j(\mathrm{fu,\ ts,\ depth})}(x_i) > 0$. Thus, each $y_i t_{j(\mathrm{fu,\ ts,\ depth})}(x_i) > 0$ defines a half space in T, and $t_{j(\mathrm{fu,\ ts,\ depth})}(x_i) = 0$ represents a hyperplane in T that acts as one of the boundaries to version space V'. This means that if we can guarantee more than $l/2$ t_j's with an accuracy of over 0.5, the result of an ensemble hypothesis might be correct.

Definition 9 (*Intersection between* $\mathbf{t}_i$ *and* t_j). $\mathrm{IntSec}(t_i, t_j) = \mathrm{Vol}(t_i) \cup \mathrm{Vol}(t_j)$.

Definition 10 (*Union between* $\mathbf{t}_i$ *and* t_j). $\mathrm{Union}(t_i, t_j) = \mathrm{Vol}(t_i) \cup \mathrm{Vol}(t_j)$.

Definition 11 (*Diversity between* $\mathbf{t}_i$ *and* t_j). $D(t_i, t_j) = \mathrm{Union}(t_i, t_j) - \mathrm{IntSec}(t_i, t_j) \approx \mathrm{Distance}(\mathrm{fu}_i, \mathrm{fu}_j) + \mathrm{Distance}(\mathrm{ts}_i, \mathrm{ts}_j) + \mathrm{Distance}(\mathrm{depth}_i, \mathrm{depth}_j)$.

Suppose the ideal hypothesis t_{ideal} and t_i might be a subset of t_{ideal}. If there are infinite t_i's, an ensemble hypothesis of them approaches the ideal hypothesis. In practice, only a finite set of t_i's is manageable. Reducing interaction between the hypotheses may lead an ensemble hypothesis approach to the ideal hypothesis rather than the others. Finally, increasing diversity results in improvement of classification performance. The following formula intuitively describes the effects of diversity.

- $\mathrm{Vol}(t_i, t_i) = \mathrm{Vol}(t_i) \leq \mathrm{Vol}(t_i, t_j)$, where t_i and t_j are different.
- $D(\mathrm{eh}_i) < D(\mathrm{eh}_j)$, where eh_i is composed of l same t_i's and eh_j is composed of t_i $(0 \leq i \leq l)$.
- $\mathrm{Vol}(\mathrm{eh}_i) = \mathrm{Vol}(t_i) \leq \mathrm{Vol}(\mathrm{eh}_j) \approx \mathrm{Vol}(t_{\mathrm{ideal}})$.

3. Diversity-based ensembling for accurate cancer classification

The proposed method consists of two parts: generating individual classification rules and combining them to construct an ensemble classifier as shown in Fig. 1. The process of generating individual classification rules is similar to approaches that have been used in previous work [41]. Feature selection is performed first to reduce the dimensionality of data, and a classification rule is generated by ensemble genetic programming. A number of individual classification rules are prepared by repeating the generation process. To combine these rules, a subset of diverse rules is constructed from the pool of individual classification rules, and then combined to produce a final decision. Conventional measures for estimating diversity are based on the output code of classifiers for training data, so diversity might depend on the characteristics of the training data. Instead, in this paper, diversity is estimated directly by comparing the structure of the rules. It does not distort the decision boundaries of base classifiers, and it does not need to concern itself with side effects from the training data.

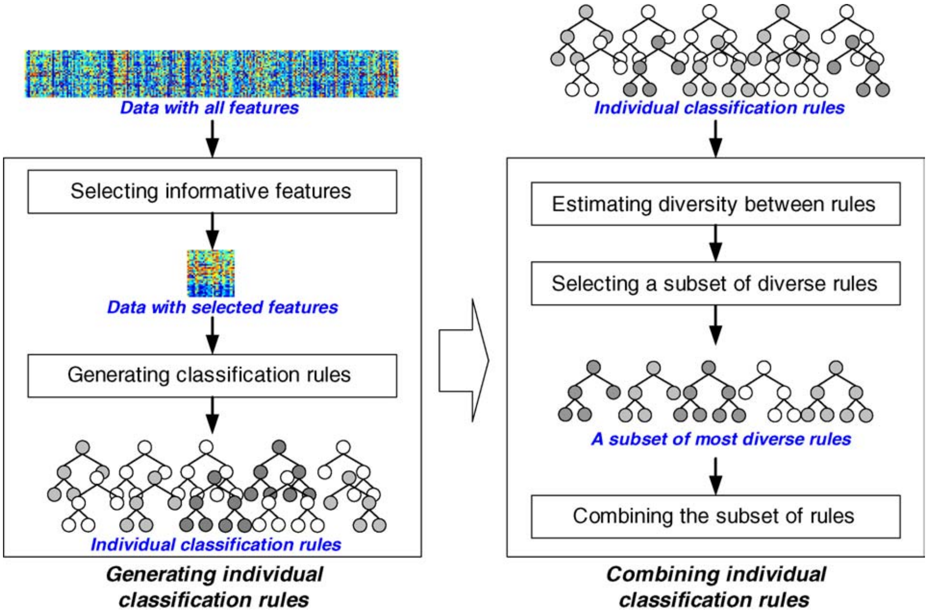

Data with all features
Selecting informative features
Data with selected features
Generating classification rules
Individual classification rules
Generating individual classification rules
Individual classification rules
Estimating diversity between rules
Selecting a subset of diverse rules
A subset of most diverse rules
Combining the subset of rules
Combining individual classification rules

Figure 1 The overview of the proposed method.

3.1. Generating multiple classification rules

DNA microarray data includes the expression information of thousands or even tens of thousands of genes, but only a small portion of them are related to the target cancer. A subset of informative genes can be selected by using the feature selection process. Cutting down the number of features to a sufficient minimum is required to improve classification performance [42].

We defined two ideal markers, obtained a standard of good features, and utilized the features by scoring the respective similarity with each ideal marker (as shown in Fig. 2). We showed that two ideal markers are negatively correlated to represent two different aspects of classification boundaries. The first marker is high in class A and low in class B, and the second marker is low in class A and high in class B. The first marker is a binary vector which consists of 1 for all the samples in class A and 0 for all the samples in class B, while the second marker is another binary vector which is composed of 0 for all the samples in class A and 1 for all the samples in class B. Five popular measures are employed. These measures are Euclidean distance, cosine coefficient, Spearman correlation, Pearson correlation, and signal-to-noise ratio [48—50]. Thirty genes are selected by each feature selection method: the first 15 for the ideal marker 1 and the rest for the ideal marker 2.

The similarity between an ideal marker ideal and a gene g can be regarded as a distance, while the distance represents how far they are located from one another. A gene is regarded as an informative gene if the distance is small, while the gene is regarded as an uncorrelated gene if the distance is large. The following formula shows the five measures used in this paper:

Pearson correlation (PC)

$$= \frac{\sum_{i=1}^{n}(\text{ideal}_i \times g_i) - \frac{\sum_{i=1}^{n}\text{ideal}_i \sum_{i=1}^{n} g_i}{n}}{\sqrt{\left(\sum_{i=1}^{n}\text{ideal}_i^2 - \frac{\left(\sum_{i=1}^{n}\text{ideal}_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} g_i^2 - \frac{\left(\sum_{i=1}^{n} g_i\right)^2}{n}\right)}}, \tag{1}$$

Spearman correlation (SC)

$$= 1 - \frac{6\sum_{i=1}^{n}(\text{ideal}_i - g_i)^2}{n \times (n^2 - 1)}, \tag{2}$$

$$\text{Euclidean distance (ED)} = \sqrt{\sum_{i=1}^{n}(\text{ideal}_i - g_i)^2}, \tag{3}$$

$$\text{Cosine coefficient (CC)} = \frac{\sum_{i=1}^{n}\text{ideal}_i \times g_i}{\sqrt{\sum_{i=1}^{n}\text{ideal}_i^2 \sum_{i=1}^{n} g_i^2}}, \tag{4}$$

$$\text{signal-to-noise ratio (SN)} = \frac{\mu_{\text{class A}}(g) - \mu_{\text{class B}}(g)}{\sigma_{\text{class A}}(g) + \sigma_{\text{class B}}(g)}, \tag{5}$$

where n is the number of samples; $\mu_{\text{class } i}(g)$ the mean of g and ideal whose label is class i; $\sigma_{\text{class } i}(g)$ the standard deviation of g; and ideal whose label is class i.

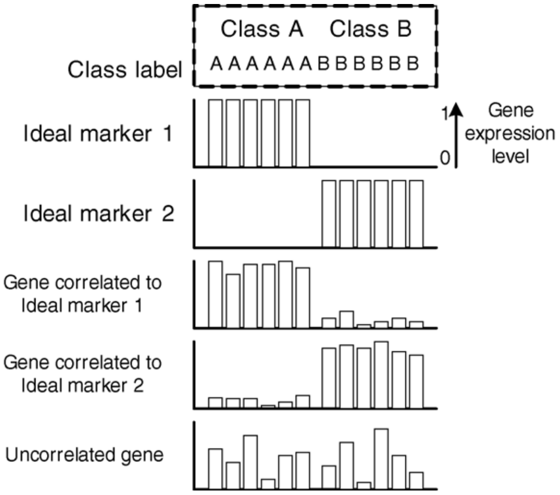


Figure 2 Negatively correlated features.

With the genes selected, genetic programming generates cancer classification rules. Arithmetic operations are employed to figure out the regulation of genes. Genetic programming uses the following procedures: initialization, evaluation, selection and genetic operation. First, the population is randomly initialized. At the evaluation stage, a fitness function evaluates individuals by estimating their fitness with regard to the target problem. Then, individuals are selected to generate the next population in proportion to the fitness. A fitter individual has more chance of being selected by the principles of natural selection and survival of the fittest. After the selection process, genetic operators such as crossover, mutation and permutation are applied to selected individuals to generate new individuals. Genetic programming repeats the process, until it finds a good solution or reaches the maximum number of generations.

In genetic programming, an individual is represented as a tree that consists of the function set {+, −, ×, ÷} and the terminal set $\{f_1, f_2, \ldots, f_n,$ constant} where n is the number of features. The function set is designed to model the up-and-down regulations of the gene expression. The classification rule is: $G = \{V = \{\text{EXP, OP, VAR}\}, T = \{+, -, \times, \div, f_1, f_2, \ldots, f_n, \text{constant}\}, P, \{\text{EXP}\}\}$, and the rule set P is as the following:

- EXP → EXP OP EXP|VAR.
- OP → + | − | × | ÷.
- VAR → $f_1 \mid f_2 \mid \ldots \mid f_n$| constant.

Classification of an instance is determined by evaluating it with the rule. An instance will be classified as class 1 if the evaluated value is larger than 0, while it will be classified as class 2 if the value is smaller than 0. Conventional genetic operators for genetic programming are employed for evolution. Crossover randomly selects and changes sub-trees from two individuals, mutation changes a sub-tree into new one, and permutation exchanges two subtrees of an individual. All genetic operations are conducted according to predefined probabilities.

A bagging approach to genetic programming generates multiple rules as shown in Fig. 1. At each evolution process, four-fifths of the training data is randomly selected to construct a training dataset. The classification accuracy of the training data is used as the fitness of an individual.

3.2. Combining diverse rules

Diversity among base classifiers is necessary to improve ensemble performance. There are various diversity measures from pair-wise to non-pair-wise such as the Q-statistic, the correlation coefficient, the Kohavi—Wolpert variance, and the entropy measure. Many researchers have studied diversity measures for improving ensemble performance. Zenobi and Cunningham used diversity based on different feature subsets [43]. Shipp and Kuncheva analyzed relationships between fusion methods and diversity measures [29], and Kuncheva and Whitaker compared various diversity measures in classifier ensembles [44]. Windeatt conducted an empirical analysis on diversity measures for the multiple classifier system [45].

When using conventional diversity approaches, there is a tradeoff (known as the accuracy—diversity dilemma) between diversity and accuracy [45]. When base classifiers have the highest levels of accuracy, diversity must decrease so that the effects of the ensemble can be reduced. The accuracy—diversity dilemma is caused by estimating the diversity of the classification results (based on the training data.) If base classifiers classify all the training data correctly, they are the same from the viewpoint of these conventional diversity measures. Moreover, it is hard to estimate diversity correctly with a few training samples like gene expression profiles. Fig. 3 shows the usefulness of the proposed method compared to the conventional diversity measure approach.

In the proposed method, a subset of diverse rules is selected in a different manner from that used in conventional approaches. The diversity among base classification rules is measured by comparing the structure of the rules. The proposed approach does not require any training data, and it is not also affected by the accuracy—diversity dilemma. Genetic programming generates an interpretable classification rule to estimate the diversity. There are several diversity measures used in genetic programming, such as pseudo-isomorphs and edit distance [46]. A simplified edit distance is used to calculate the diversity between the classification rules, and the distance between two rules r_i and r_j is estimated as follows:

$$\text{distance}(r_i, r_j) = \begin{cases} d(p, q) \\ \text{if neither} r_i \text{ nor} r_j \text{ have any children}, \\ d(p, q) + \text{distance}(\text{RS of} r_i, \\ \text{RS of} r_j) + \text{distance}(\text{LS of} r_i, \text{LS of} r_j) \\ \text{otherwise}(\text{RS : right subtree, LS :} \\ \text{left subtree}) \end{cases},$$

$$\text{where } d(p, q) = \begin{cases} 1 & \text{if } p \text{ and} q \text{ overlap} \\ 0 & \text{if } p \text{ and} q \text{ do not overlap} \end{cases}$$

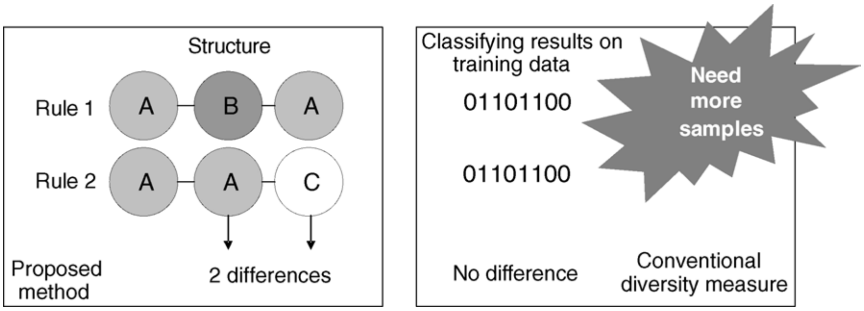


Figure 3 A comparison between the proposed method and the conventional diversity measure.

The appearance of genes in the rules is also used to measure diversity. Diversity decreases when two rules use a same gene, while it increases if the rules are different genes. Finally, five classification rules are selected to compose an ensemble classifier as follows:

```
$R$: A set of extracted rules $\{r_1, r_2, \ldots, r_n\}$
$S$: A set of selected rules $\{s_1, s_2, \ldots, s_m\}$
int calculate_diversity($r_i$, $r_j$) {
  $cf_{ij}$ = common_feature_number($r_i$, $r_j$);
  $df_{ij}$ = different_feature_number($r_i$, $r_j$);
  $ed_{ij}$ = distance($r_i$, $r_j$);
  return $df_{ij} - cf_{ij} - \alpha \times ed_{ij}$;
}
For $i = 1$ to $n$ {
  For $j = i + 1$ to $n$ {
    $d_{ij}$ = calculate_diversity($r_i$, $r_j$);
}}
Find a set $S$ in which rules' diversity is maximized
$S = \{s_1, s_2, \ldots, s_m\}$
```

A fusion method combines these rules to generate the final classification results. Five simple combination methods that do not require training are majority vote (MAJ), maximum (MAX), minimum (MIN), average (AVG) and product (PRO). Three sophisticated combination methods that do require training are Naïve Bayes (NB), behavior-knowledge space (BKS), and decision templates (DT). Oracle (ORA) is only used to show a possible upper limit to classification accuracy. Detailed explanations of the fusion methods can be found in [23].

4. Experimental results

4.1. Experimental environment

There are several DNA microarray datasets from published cancer gene expression studies. These include breast cancer datasets, central nervous system cancer datasets, colon cancer datasets, leukemia cancer datasets, lung cancer datasets, lymphoma cancer datasets, NCI60 datasets, ovarian cancer datasets, and prostate cancer datasets. Among them, three representative datasets were used in this paper. The first and second datasets involve samples from two variants of the same disease and the third involves a tumor and normal samples of the same tissue.

- *Lymphoma cancer dataset* [47]: Diffuse large B-cell lymphoma (DLBCL) is a disease, which is the common sub-type of non-Hodgkin's lymphoma. There are various subtypes of lymphoma cancer that require different treatment, but it is not easy to distinguish them clinically. Hence, the lymphoma cancer classification using gene expression profiles has been investigated [48]. The gene expression data (http://llmpp.nih.gov/lymphoma/; accessed: 30 April 2005) contains 4026 genes across 47 samples: 24 samples of the germinal center B-like group and 23 samples of the activated B-like group.
- *Lung cancer dataset* [51]: This gene expression data has been used to classify malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissues: 31 MPM tissues and 150 ADCA tissues, while each tissue has 12,533 gene expression levels.
- *Ovarian cancer dataset* [52]: This gene expression data aims to identify proteomic patterns in serum to distinguish ovarian cancer. It has 91 controls (normal) and 162 ovarian cancer tissues. Each sample has 15,154 gene expression levels.

Each feature selection method scores genes, and selects 30 top-ranked genes as the feature of the input pattern. For classification, genetic programming is set (as shown in Table 1). We conducted a five-fold cross-validation for each dataset. In five-fold cross-validation, one-fifth of all samples is evaluated as test data while the others are used as training data. The training data is used to select informative features. This process is repeated 100 times to obtain the average results with 500 (5 × 100) experiments in total.

4.2. Results on classification accuracy

Tables 2—4 summarize the predictive accuracy of the proposed method for each cancer dataset; the highlighted values represent high accuracy. '10 classifiers' and '5 classifiers' are based on ensembling

Table 1 Experimental environments

Parameter	Value	Parameter	Value
Population size	200	Mutation rate	0.1—0.3
Maximum generation	3000	Permutation rate	0.1
Selection rate	0.6—0.8	Maximum depth of a tree	3—5
Crossover rate	0.6—0.8	Elitism	Yes

Table 2 Test accuracy on lymphoma cancer dataset (%)

Features	Fusion method	10 classifiers	5 classifiers	5 diverse classifiers	Individual classifier
PC	MAJ	95.0	93.2	97.1	91.3
	MAX	96.7	95.2	97.6	
	MIN	81.7	82.5	83.7	
	AVG	96.7	94.4	96.9	
	PRO	80.5	72.0	78.1	
	NB	65.7	72.5	70.7	
	BKS	95.0	93.2	97.1	
	DT	95.8	95.1	97.0	
	ORA	100	100	100	
SC	MAJ	90.7	91.0	92.1	88.1
	MAX	91.9	89.7	93.9	
	MIN	81.4	80.2	79.0	
	AVG	91.8	89.4	94.2	
	PRO	84.7	82.4	81.0	
	NB	61.2	66.7	67.4	
	BKS	90.7	91.0	92.1	
	DT	92.2	88.6	94.4	
	ORA	100	98.4	99.3	
ED	MAJ	91.6	93.3	95.6	88.2
	MAX	92.2	90.1	93.2	
	MIN	77.7	81.8	80.2	
	AVG	93.4	92.7	95.2	
	PRO	80.9	79.0	78.5	
	NB	60.8	69.2	65.8	
	BKS	91.6	93.3	95.6	
	DT	93.7	92.6	94.3	
	ORA	98.4	98.4	97.8	
CC	MAJ	91.6	93.3	94.1	88.8
	MAX	90.6	90.6	91.6	
	MIN	80.1	82.9	81.8	
	AVG	91.4	92.9	92.6	
	PRO	84.5	79.4	77.6	
	NB	63.6	74.7	70.5	
	BKS	91.6	93.3	94.1	
	DT	92.9	93.0	93.1	
	ORA	100	98.4	100	
SN	MAJ	94.5	94.8	94.8	90.4
	MAX	97.6	94.7	96.5	
	MIN	81.9	83.7	80.8	
	AVG	96.0	95.6	96.2	
	PRO	88.0	83.9	84.9	
	NB	60.6	72.0	69.0	
	BKS	94.5	94.8	94.8	
	DT	96.0	94.2	95.0	
	ORA	100	100	100	

from random forest, while '5 diverse classifiers' is the result of the proposed method. The results show that the ensemble classifier performs better than the individual classifier. In most cases, 1—10% increments are observed when using ensemble techniques. In addition, the proposed method shows superior classification performance when combining the 10 rules and the 5 rules. It signifies that considering diversity improves the performance of the ensemble.

The ensemble that uses 10 classification rules is sometimes inferior to the ensemble that uses 5 classification rules, even though the former procedure includes more information than the latter.

Table 3 Test accuracy on lung cancer dataset (%)

Features	Fusion method	10 classifiers	5 classifiers	5 diverse classifiers	Individual classifier
PC	MAJ	98.9	99.1	99.2	98.2
	MAX	99.5	98.9	99.4	
	MIN	96.0	96.7	96.7	
	AVG	99.4	99.3	99.4	
	PRO	96.5	95.1	94.6	
	NB	93.7	95.5	95.4	
	BKS	98.9	99.1	99.2	
	DT	99.1	99.0	99.2	
	ORA	100	99.8	99.8	
SC	MAJ	98.8	98.8	99.0	97.8
	MAX	99.3	99.2	99.4	
	MIN	94.3	95.9	96.1	
	AVG	99.3	99.2	99.4	
	PRO	96.3	95.4	94.4	
	NB	91.4	94.2	94.2	
	BKS	98.8	98.8	99.0	
	DT	99.0	98.9	99.1	
	ORA	99.8	99.7	99.6	
ED	MAJ	94.7	94.8	95.1	94.6
	MAX	94.8	94.6	94.3	
	MIN	94.8	94.8	94.4	
	AVG	94.4	94.6	94.6	
	PRO	92.8	92.7	93.4	
	NB	95.4	95.3	94.9	
	BKS	95.6	95.3	95.8	
	DT	94.6	94.1	94.3	
	ORA	97.2	96.7	96.6	
CC	MAJ	99.4	99.3	99.3	98.5
	MAX	99.3	99.2	99.3	
	MIN	97.2	97.3	96.9	
	AVG	99.4	99.3	99.4	
	PRO	97.7	96.8	97.1	
	NB	95.7	96.6	96.3	
	BKS	99.4	99.3	99.2	
	DT	99.1	98.7	98.7	
	ORA	99.6	99.5	99.6	
SN	MAJ	99.1	98.9	99.3	98.0
	MAX	99.3	99.3	99.4	
	MIN	94.8	96.3	96.1	
	AVG	99.4	99.2	99.3	
	PRO	96.5	95.0	95.0	
	NB	91.4	94.6	94.3	
	BKS	99.1	98.9	99.3	
	DT	99.2	99.1	99.2	
	ORA	99.9	99.8	99.8	

This implies that error is increased with increasing number of base classifiers. Theoretical proof of this can be found in [28]. In this experiment, however, most cases show that the 10-classifier ensemble is better than the 5-classifier ensemble. Finally, the proposed approach not only supports the same degree of useful information with the ensemble that uses 10 classification rules, but also minimizes the increment of the error.

In fusion methods, MAX and AVG distinguish themselves by combining multiple classification rules obtained by genetic programming, while BKS and DT also show good performance. Since the proposed method classifies samples with a linear boundary

Table 4 Test accuracy on ovarian cancer dataset (%)

Features	Fusion method	10 classifiers	5 classifiers	5 diverse classifiers	Individual classifier
PC	MAJ	97.0	97.2	97.1	96.9
	MAX	97.4	97.4	97.4	
	MIN	95.1	95.8	95.6	
	AVG	97.4	97.3	97.5	
	PRO	94.2	93.9	94.0	
	NB	96.2	96.7	96.4	
	BKS	97.0	97.1	97.1	
	DT	97.8	98.0	98.0	
	ORA	98.3	97.9	98.2	
SC	MAJ	97.1	97.1	96.9	96.9
	MAX	97.2	97.1	97.3	
	MIN	95.7	95.9	95.9	
	AVG	97.2	97.1	97.3	
	PRO	95.5	95.5	95.2	
	NB	97.2	97.1	97.3	
	BKS	97.1	97.1	97.0	
	DT	97.6	97.5	97.5	
	ORA	97.9	97.7	97.9	
ED	MAJ	93.8	93.9	94.3	94.2
	MAX	94.9	94.7	95.1	
	MIN	93.7	93.8	93.8	
	AVG	94.3	94.5	95.0	
	PRO	86.7	86.9	86.9	
	NB	94.3	94.3	94.5	
	BKS	93.8	93.9	94.0	
	DT	94.3	94.3	94.4	
	ORA	95.8	95.4	95.8	
CC	MAJ	84.6	84.8	84.9	84.4
	MAX	85.4	85.3	85.4	
	MIN	82.8	83.4	82.8	
	AVG	85.4	85.4	85.4	
	PRO	75.7	75.5	75.0	
	NB	84.5	84.6	84.5	
	BKS	84.4	84.6	84.9	
	DT	82.9	82.1	81.9	
	ORA	86.5	86.3	86.5	
SN	MAJ	97.0	96.2	97.0	96.7
	MAX	97.2	96.2	97.2	
	MIN	94.9	92.4	95.5	
	AVG	97.2	97.2	97.3	
	PRO	94.9	93.7	93.9	
	NB	95.9	96.4	96.2	
	BKS	97.0	97.0	97.0	
	DT	97.7	97.6	97.8	
	ORA	97.6	97.5	97.6	

obtained by genetic programming, the margin affects the performance of these fusion methods. MAX selects a rule with the maximum margin, while AVG sums up the margins of all rules. However, MIN, PRO and NB work poorly in classification, since their approach is to minimize the risk. If there is poor classification among the pool of rules, the overall performance of the ensemble might decline. Especially, MAX, AVG, PRO and DTare improved when the ensemble combines more classification rules. This signifies that these fusion methods are dependent on the number of classifiers, so it might be helpful to generate more classifiers when using the fusion methods.

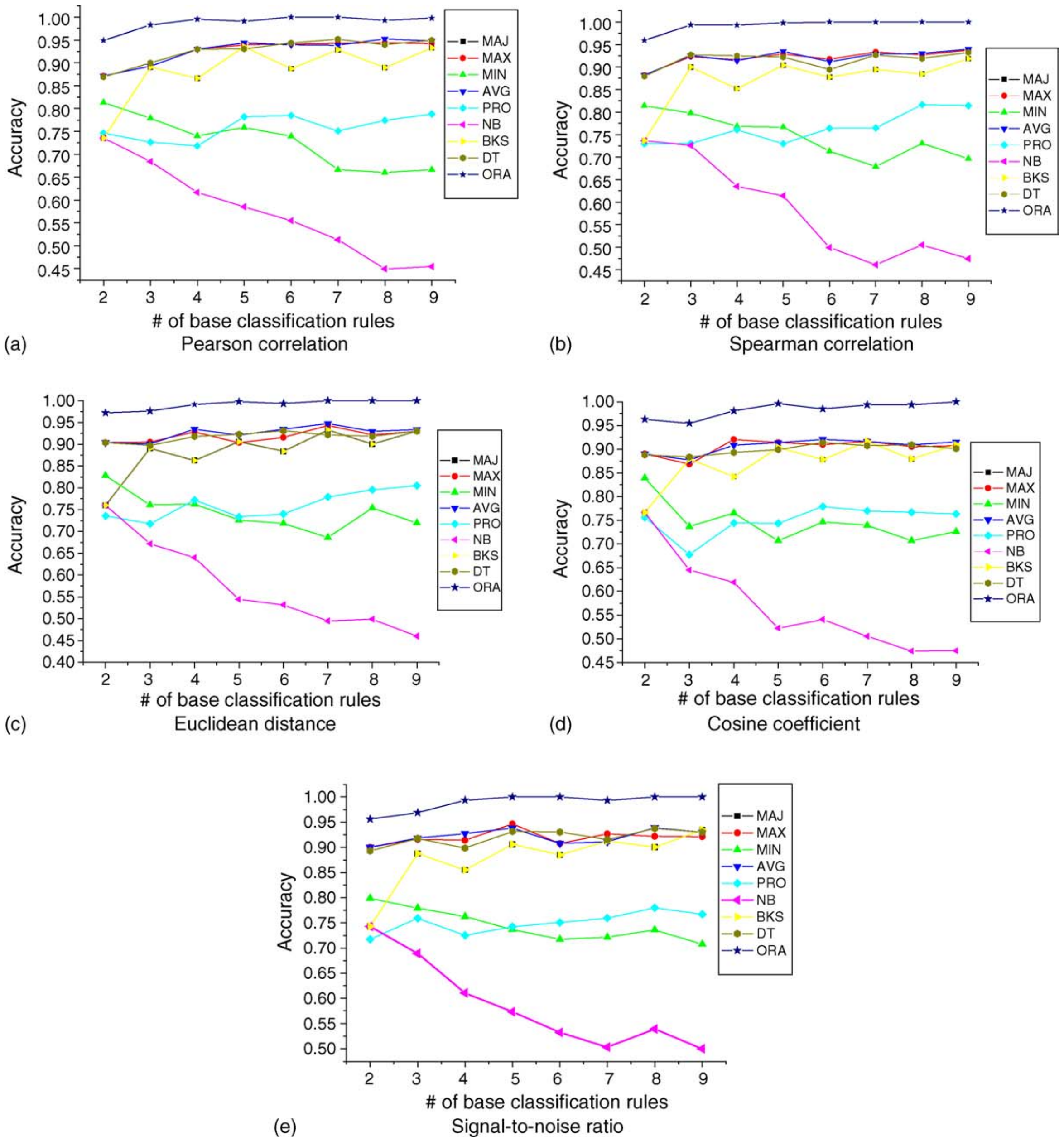


Figure 4 Test accuracy for lymphoma cancer dataset according to the number of base classification rules: (a) Pearson correlation; (b) Spearman correlation; (c) Euclidean distance; (d) cosine coefficient; (e) signal-to-noise ratio.

The relationship between the number of base classification rules and the performance of the ensemble is examined and shown in Fig. 4. Performance increases with the number of base classification rules, but it is almost converged when using four to six rules. BKS shows an oscillation between even and odd numbers. When the ensemble uses the odd number of rules, it shows better performance than when using even numbers. The accuracy of MIN and NB is decreased with the increment of the number of base classification rules, since they are subject to noise. Even though PRO is poorer than the individual classification rule, the accuracy of PRO is gradually increased with the addition of base classification rules. Especially, the performance of MIN and NB falls down, since they are sensitive to noise.

The performance of the features for three datasets is compared in Fig. 5. The PC and the SN work better than the others, while the SC shows a wide variance of classification. The PC shows good performance for the lymphoma cancer dataset. The PC, SC and SN work well for the lung cancer dataset, and the SC and SN obtain a high accuracy for the ovarian cancer dataset. We can find out a correlation

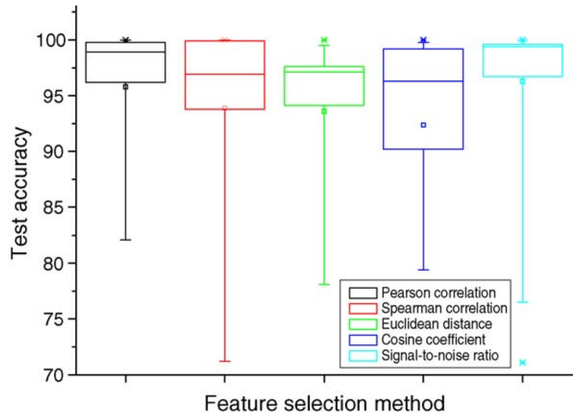
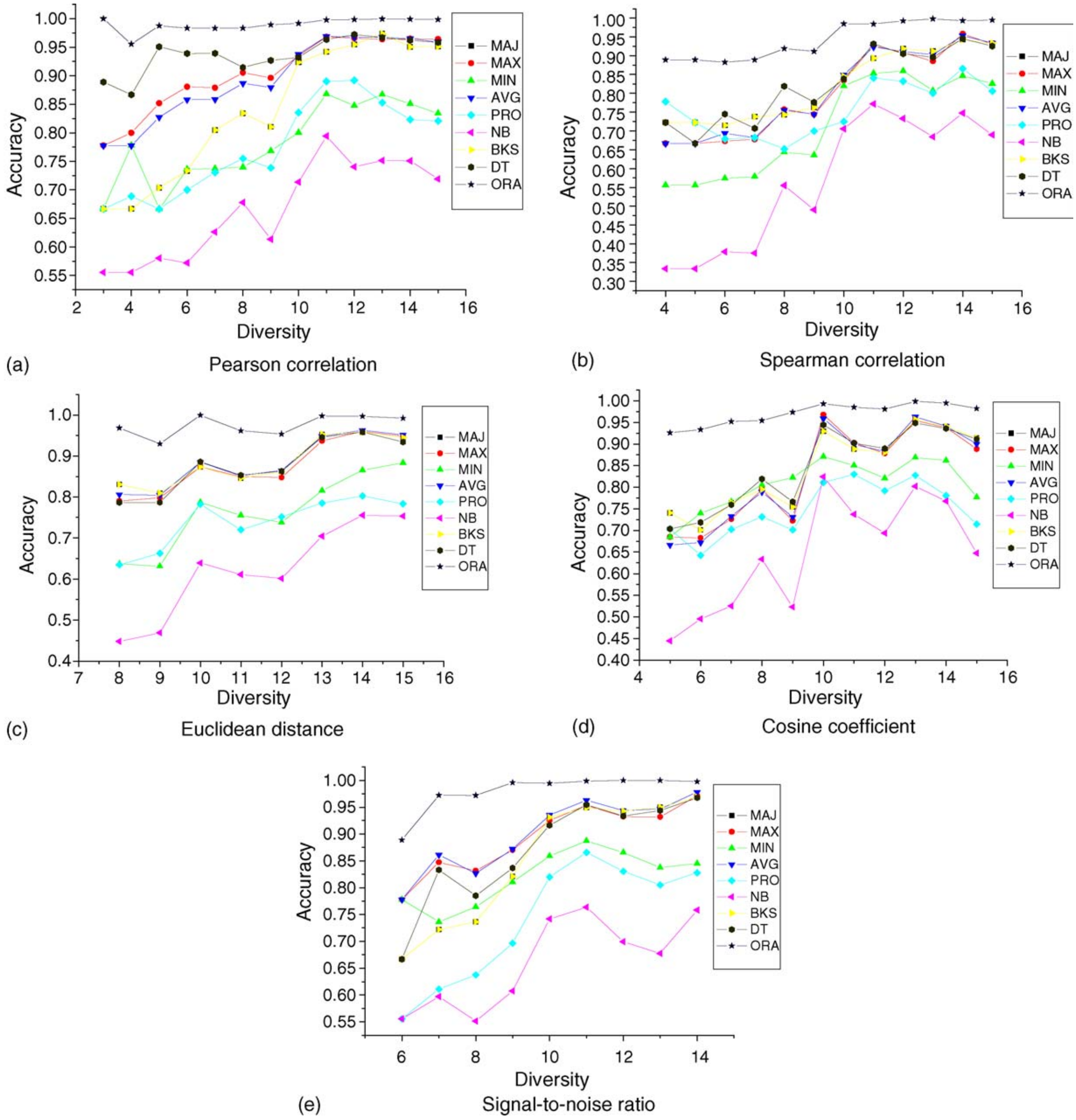


Figure 5 Test accuracy for the feature selection method.

method so that the PC, SC and SN might work better than a simple distance measurement such as ED and CC for the three datasets.

4.3. Results on performance by diversity

The relationship between diversity and performance is also analyzed and shown in Fig. 6. The results indicate that classification accuracy increases according to the increment of diversity in most cases. A decline in accuracy occasionally appears, because diversity is apt to increase when there is a peculiar rule. This can be solved



(a) Pearson correlation

(b) Spearman correlation

(c) Euclidean distance

(d) Cosine coefficient

(e) Signal-to-noise ratio

Figure 6 The predictive accuracy according to diversity for Lymphoma cancer dataset: (a) Pearson correlation; (b) Spearman correlation; (c) Euclidean distance; (d) cosine coefficient; (e) signal-to-noise ratio.

by a non-pair-wise approach for estimating diversity in ensemble genetic programming.

Experiments on conventional diversity measurements such as the *Q*-statistics, the correlation coefficient, the disagreement measure, the double-fault measure, the Kohavi—Wolpert variance, measurement of interrater agreement, the entropy measure, the measure of difficulty, generalized diversity, and coincident failure diversity were conducted to compare with the proposed method. These diversity measurements are explained in [29,44].

As mentioned before, conventional diversity measurements have limitations in applying cancer classification when using gene expression data. They require a number of samples to measure diversity correctly, and there is less enhancement when base classifiers are highly accurate. Most gene expression datasets provide only a few samples, and the classification rules obtained by genetic programming produce high accuracy (as shown in Tables 2—4). In most cases, the individual classification rules obtain 100% accuracy for training data, so conventional diversity measurements output only 0 for all

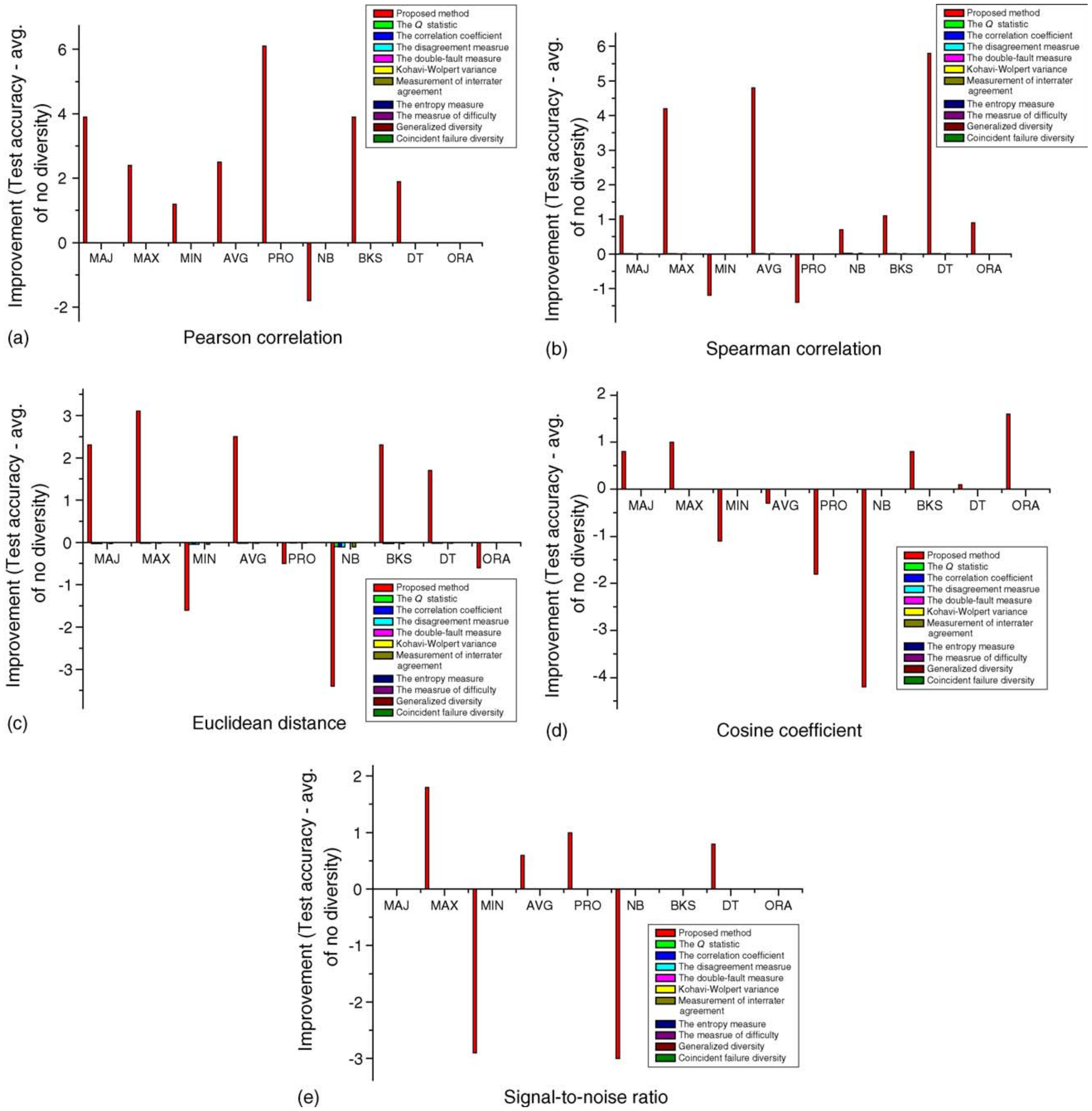


Figure 7 The improvement of accuracy according to the diversity measurement for the lymphoma cancer dataset: (a) Pearson correlation; (b) Spearman correlation; (c) Euclidean distance; (d) cosine coefficient; (e) signal-to-noise ratio.

Table 5 Comparison on lymphoma cancer dataset (%)

Features	Fusion method	Ensemble neural networks (5 classifiers)	5 diverse classifiers
PC	MAJ	93.1	97.1
	MAX	92.8	97.6
	AVG	93.1	96.9
	BKS	93.1	97.1
	DT	93.2	97.0
	ORA	94.2	100
SC	MAJ	93.3	92.1
	MAX	92.9	93.9
	AVG	93.3	94.2
	BKS	93.3	92.1
	DT	92.9	94.4
	ORA	93.3	99.3
ED	MAJ	92.8	95.6
	MAX	93.3	93.2
	AVG	92.8	95.2
	BKS	92.8	95.6
	DT	92.6	94.3
	ORA	97.1	97.8
CC	MAJ	93.3	94.1
	MAX	92.6	91.6
	AVG	93.3	92.6
	BKS	93.3	94.1
	DT	92.9	93.1
	ORA	94.2	100
SN	MAJ	93.3	94.8
	MAX	93.6	96.5
	AVG	93.3	96.2
	BKS	93.3	94.8
	DT	93.4	95.0
	ORA	94.2	100

possible combinations of classification rules. Instead, since the proposed method does not consider the training data, it produces improved performance in the ensemble. Fig. 7 shows the proposed method compared to conventional diversity measurements. In most cases, classification rules classify all training data correctly. The proposed method obtained higher accuracy than the ensemble without diversity while the other diversity measurements did not provide any improved accuracy.

4.4. Comparison with the neural networks

In order to compare with the other representative machine learning technique, we classified the lymphoma cancer dataset using the neural networks. As similar to the proposed method, five neural networks are combined to generate a final output. Table 5 shows the result of the experiment, and the proposed method produced better performance than the neural networks in most cases.

5. Conclusion

The classification of cancer, based on gene expression profiles, is a challenging task in bioinformatics. Many machine-learning techniques have been developed to obtain highly accurate classification performance. In this paper, we have proposed an effective ensemble approach that uses diversity in ensemble genetic programming to classify gene expression data. The ensemble helps improve classification performance, but diversity is also an important factor in constructing an ensemble classifier. Contrary to conventional diversity-based ensemble approaches, diversity in this paper was measured by comparing the structure of classification rules. The proposed method is independent from the training data, so that it can be effective in cancer classification using gene expression data with only a few training samples. Experiments on three representative DNA microarray datasets have demonstrated the usefulness of the proposed method. Moreover, several rank-based feature selection methods and fusion methods were also compared. The SC, PC and SN showed good performance in classifying the datasets while the MAX, AVG, BKS and DT effectively combined base classification rules so as to obtain high accuracies.

In this work, a simple distance measurement was used to estimate diversity among classification rules, so there was a limitation when calculating the correct diversity. In future work, a more sophisticated distance measurement for ensemble genetic programming will be developed for measuring accurate diversity. Moreover, a non-pair-wise approach will be also studied for estimating diversity.

Acknowledgement

This research was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Commerce, Industry and Energy.

References

[1] Schmidt U, Begley C. Cancer diagnosis and microarrays. Int J Biochem Cell Biol 2003;35(2):119—24.

[2] Lu Y, Han J. Cancer classification using gene expression data. Inform Syst 2003;28(4):243—68.

[3] Sarkar I, Planet P, Bael T, Stanley S, Siddall M, DeSalle R, et al. Characteristic attributes in cancer microarrays. J Biomed Inform 2002;35(2):111—22.

[4] Kuo W, Kim E, Trimarchi J, Jenssen T, Vinterbo S, Ohno-Machado L. A primer on gene expression and microarrays for machine learning researchers. J Biomed Inform 2004;37(4): 293—303.

[5] Azuaje F. A computational neural approach to support the discovery of gene function and classes of cancer. IEEE Trans Biomed Eng 2001;48(3):332—9.

[6] Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):673—9.

[7] Albrecht A, Vinterbo S, Ohno-Machado L. An epicurean learning approach to gene-expression data classification. Artif Intell Med 2003;28(1):75—87.

[8] Huang C, Liao W. Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. Neural Process Lett 2004;19(3):211—26.

[9] Roth V, Lange T. Bayesian class discovery in microarray datasets. IEEE Trans Biomed Eng 2004;51(5):707—18.

[10] Zhou X, Liu K, Wong S. Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform 2004;37(4):249—59.

[11] Pan F, Wang B, Hu X, Perrizo W. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis. J Biomed Inform 2004;37(4):240—8.

[12] Ding C, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17(4):349—58.

[13] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci 2001;98(26): 15149—54.

[14] Camp N, Slattery M. Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer. Cancer Causes Contr 2002;13(9): 813—23.

[15] Zhang H, Yu C, Singer B. Cell and tumor classification using gene expression data: construction of forests. Proc Natl Acad Sci 2003;100(7):4168—72.

[16] Li L, Weinberg C, Darden T, Pedersen L. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17(12):1131—42.

[17] Deutsch J. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics 2003;19(1): 45—52.

[18] Karzynski M, Mateos A, Herrero J, Dopazo J. Using a genetic algorithm and a perceptron for feature selection and supervised class learning in DNA microarray data. Artif Intell Rev 2003;20(1/2):39—51.

[19] Langdon W, Buxton B. Genetic programming for mining DNA chip data for cancer patients. Genet Programm Evol Mach 2004;5(3):251—7.

[20] Valentini G. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. Artif Intell Med 2002;26(3):281—304.

[21] Park C, Cho S-B. Evolutionary computation for optimal ensemble classifier in lymphoma cancer classification. In: Lecture notes in artificial intelligence (ISMIS), vol. 2871; 2003. p. 521—30.

[22] Tan A, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinform 2003; 2(Suppl 3):S75—83.

[23] Kuncheva L. A theoretical study on six classifier fusion strategies. IEEE Trans Pattern Anal Mach Intell 2002;24(2): 281—6.

[24] Bryll R, Guttierez-Osuna R, Quek F. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recog 2003;36(6):1291—302.

[25] Webb G, Zheng Z. Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. IEEE Trans Knowledge Data Eng 2004;16(8):980—91.

[26] Optiz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999;11:169—98.

[27] Islam M, Yao X, Murase K. A constructive algorithm for training cooperative neural network ensembles. IEEE Trans Neural Network 2003;14(4):820—34.

[28] Zhou Z, Wu J, Tang W. Ensemble neural networks: many could be better than all. Artif Intell 2002;137(1/2):239—63.

[29] Shipp C, Kuncheva L. Relationships between combination methods and measures of diversity in combining classifiers. Inform Fusion 2002;3(2):135—48.

[30] Verikas A, Lipnickas A, Malmqvist K, Bacauskiene M, Gelzinis A. Soft combination of neural classifiers: a comparative study. Pattern Recog Lett 1999;20(4):429—44.

[31] Tax D, Van Breukelen M, Duin R, Kittler J. Combining multiple classifiers by averaging or by multiplying? Pattern Recog 2000;33(9):1475—85.

[32] Alexandre L, Campihlo A, Kamel M. On combining classifiers using sum and product rules. Pattern Recog Lett 2001; 22(12):1283—9.

[33] Kuncheva L, Bezdek J, Duin K. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recog 2001;34(2):299—314.

[34] Kim K, Cho S-B. Fuzzy integration of structure adaptive SOMs for web content mining. Fuzzy Sets Syst 2004;148(1):43—60.

[35] Koza J. Genetic programming. In: Williams J, Kent A, editors. Encyclopedia of computer science and technology, vol. 39. New York: Marcel Dekker; 1998. p. 29—43.

[36] Zhang Y, Bhattacharyya S. Genetic programming in classifying large-scale data: an ensemble method. Inform Sci 2004;163(1—3):85—101.

[37] Brameier M, Banzhaf W. Evolving teams of predictors with linear genetic programming. Genet Programm Evol Mach 2001;2(4):381—407.

[38] Fernandez F, Tomassini M, Vanneschi L. An empirical study of multipopulation genetic programming. Genet Programm Evol Mach 2003;4(1):21—51.

[39] Imamura K, Soule T, Heckendorn R, Foster J. Behavioral diversity and a probabilistically optimal GP ensemble. Genet Programm Evol Mach 2003;4(3):235—53.

[40] Tong S, Koller D. Support vector machine active learning with applications to text classification. J Mach Learning Res 2001;2:45—66.

[41] Hong J-H, Cho S-B. Lymphoma cancer classification using genetic programming with SNR features. In: Lecture notes in computer science (EuroGP), vol. 3003; 2004. p. 78—88.

[42] Xiong M, Li W, Zhao J, Jin L, Boerwinkle E. Feature selection in gene expression-based tumor classification. Mol Genet Metab 2001;73(3):239—47.

[43] Zenobi G, Cunningham P. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In: Lecture Notes in Computer Science (ECML), vol. 2167; 2001. p. 576—87.

[44] Kuncheva L, Whitaker C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learning 2003;51(2):181—207.

[45] Windeatt T. Diversity measures for multiple classifier system analysis and design. Inform Fusion 2005;6(1):21—36.

[46] Burke E, Gustafson S, Kendall G. Diversity in genetic programming: an analysis of measures and correlation with fitness. IEEE Trans Evol Comput 2004;8(1):47—62.

[47] Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403(6769):503—11.

[48] Shipp M, Ross K, Tamayo P, Weng A, Kutok J, Aguiar R, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8(1):68—74.

[49] Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531—7.

[50] Radonic A, et al. Reference gene selection for quantitative real-time PCR analysis in virus infected cells: SARS corona virus, Yellow fever virus, Human Herpesvirus-6, Camelpox virus and Cytomegalovirus infections. Virol J 2005;2(1):7.

[51] Gordon G, Jensen R, Hsiao L, Gullans S, Blumenstock J, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62(17):4963—7.

[52] Petricoin E, Ardekani A, Hitt B, Levine P, Fusaro V, Steinberg S, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359(9306):572—7.