International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.6 (2015), pp.203-218
http://dx.doi.org/10.14257/ijmue.2015.10.6.21

A Novel Selective Ensemble Classification of Microarray Data Based on Teaching-Learning-Based Optimization

Tao Chen[1, 2*], Zenglin Hong[1], Fang-an Deng[2], Xiao Yang[2], Jun Wei[2] and Man Cui[1]

1 School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China
2 School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, Shaanxi, 723000, China
**Corresponding author:ct79hz@126.com*

Abstract

Aiming at the characteristics of high dimension and small samples in microarray data, this paper proposes a selective ensemble method to classify microarray data. Firstly, kruskal-wallis test is used to filter irrelevant genes with classification task and to obtain a set of genes, and then a reduced training set is produced from original training set according to gene subset obtained. Secondly, multiple gene subsets are generated by using neighborhood rough set model with different radius and used to construct training subsets on above reduced training set. Thirdly, every constructed training subset is used to train a classifier by using SVM algorithm, and then multiple classifiers are produced as base classifiers. Finally, a set of base classifiers are selected by using teaching-learning-based optimization and build an ensemble classifier by weighted voting. Five benchmarks tumor microarray datasets are applied to evaluate performance of our proposed method. Experimental results indicate our proposed method is very effective and efficient for classifying microarray data, and it improves not only classification accuracy, but also decrease memory costs and computation times.

Keywords: *DNA microarray; selective ensemble classification; kruskal-wallis test; neighborhood rough set model; teaching-learning-based optimization*

1. Introduction

The considerable knowledge of tumor-related molecular biomedicine indicates that the tumor is recognized as a complex systems biology disease since its genesis and development involves the complicated spatiotemporal organization of signaling pathway. In the past, tumor diagnosis depends on using a complex combination of clinical and histopathological data. However, it is often difficult or impossible to recognize tumor types in typical instances. With the development of DNA microarray technology, it is possible to detect the expression levels of thousands of genes in a single experiment, and it will help to classify diseases according to expression levels in normal and tumor cells from molecular biology [1-4].Therefore, DNA microarray data classification is attracting more and more attention and research.

Up to now, many machine learning methods, such as decision tree [5], artificial neural network (ANN) [6,7], bayesian networks [8], k-nearest neighbor (KNN) [9,10] and support vector machine (SVM) [11-14], *etc.*, were utilized to classify microarray data and have obtained certain success. However, the characteristic of high dimension and small samples in microarray data lead to lower performance of these classification methods. Especially, different classification methods can also

ISSN: 1975-0080 IJMUE

achieve different results for certain special problem, and then it leads to increase the risk of selecting the classification methods of poor performance. In order to solve this problem, ensemble classification was proposed and applied to classify microarray data, which is to train multiple bases classifiers and combine outputs to classify new samples. The ensemble classification is adopted to improve the overall classification performance, because the errors of one classifier are averaged out by the correct classification of another classifier. Therefore, ensemble learning can decrease the risk of selecting a poor performance classifier and usually improve classification performance [15].

Many researches indicate diversity and accuracy of bases classifiers are important to affect ensemble classification performance. The bigger diversity and higher accuracy can improve ensemble classification performance. Feature disturbance is an effective method for increasing diversity among bases classifiers, such as random sunspace [16], random forest [17], *etc*. Neighborhood rough set model [18,19], proposed by Hu qinghua, is an improved attributes reduction method on the basis of rough set theory, which can directly deal with continuous attributes to avoid information loss. Research shows neighborhood rough set model is an effective attribute reduction method than rough set. In neighborhood rough set model, radius of neighborhood is important factor to affect attribute reduction performance and different attribute subsets are produced by using neighborhood rough set model with different radius and have high diversity. Therefore, this paper applies neighborhood rough set model with different radius to generate different attribute subsets, and then multiple training subsets are produced according to above different attribute subsets. It guarantees the high diversity among base classifiers because of the diversity among training subsets.

At present, most ensemble methods combine outputs of all base classifiers, and it leads to increase of computation time and storage space. Moreover, it does not always improve classification performance. Selective ensemble is effective for improving ensemble performance and decreasing computation time and storage space, which is to select a set of base classifiers to combine output [20].

Teaching-Learning-Based optimization (TLBO), proposed by R. V. Rao in 2011, is a novel intelligent optimization algorithm based on population search [21]. TLBO simulates teaching behavior of teacher and learning behavior of learners in a class to improve student achievement. Many optimization algorithms, such as genetic algorithm (GA), particle swarm optimization (PSO) and harmony Search (HS) were proposed and widely used in optimization problems. However, these optimization algorithms need to set algorithm parameters in advance, which highly affect the optimization performance, moreover unsuitable parameters usually reduce the performance of optimization algorithm. It is difficult that parameters of optimization algorithm is set correctly, and it leads to a limitation for the widespread application of the optimization algorithms. However, any parameters need not be set in advance for TLBO. Therefore, TLBO is widely used in optimization problems because of no parameters, simple principle, fast speed, high precision and better overall search ability to compare with the others [22-23].

This paper proposes a selective ensemble method to classify microarray data. Firstly, genes are ranked by using kruskal-wallis test method to preselect genes subset from original gene set, and then training set is reduced to construct new training set; Secondly, multiple gene subsets are obtained from above preselected gene subset based on neighborhood rough set model with different radius, and then multiple training subsets are generated from training set reduced according to different gene subsets. Here, the diversity among training subsets obtained is more large because of the diversity among gene subsets. Thirdly, every training subset is applied to train a base classifier, and then multiple base classifiers are produced.

Finally, a set of base classifiers are selected by using teaching-learning-based optimization and combined to build an ensemble classifier by weighted voting.

The remainder of this paper is organized as follows: Section 2 introduces basic ideas and steps about methods, including to kruskal-wallis test, neighborhood rough set model and Teaching-Learning-Based optimization. In Section 3, a selective ensemble classification method based on neighborhood rough set model and teaching-learning-based optimization is proposed, and ideas and flow chart of our proposed method are given. Section 4 makes experiment on five benchmarks tumor microarray datasets and gives the experimental results and analysis. The conclusion is made in Section 5.

2. Materials and Methods

2.1 Kruskal-Wallis Test

The kruskal-wallis test is a non-parametric alternative to the well-known one-way independent samples analysis of variance [24]. The null hypothesis of the test is that the samples come from populations with equal medians. Given n_c groups, the kruskal-wallis test statistic should be compared with the χ^2 statistic with $n_c - 1$ degrees of freedom if the sample size within each group is large enough (*e.g.*, >5). This score is derived for all the features so they can be ranked according to their χ^2 value. The different models are built by removing the variables with the smallest χ^2 value. In the end, the variables that are included in the model best performing on validation data, using stratified random sampling, are selected for use on test data. This procedure selects optimal variables in a relatively fast way without causing a massive search process.

2.2 Neighborhood Rough Set Model

Rough set (RS) is a math analysis tool and was brought up by Pawlak in 1982 to effectively process incomplete and inaccurate information. Rough set don't need any prior information and can only rely on internal information of data themselves to discover tacit knowledge within them, reveal potential rules and effectively process incomplete and inaccurate data. In traditional rough set, continuous data must be first discretized, which will result in original information loss, and the results of calculation and process are highly affected by discretization method. Neighborhood rough set model, brought up by Hu qinghua, is an improved method that develops from classical rough set and can directly process continuous data [18,19]. It needn't discretize continuous data in advance, and can be directly used for problems of knowledge reduction.

For discrete data, the samples with the same feature value are pooled into a set, called equivalence class. These samples are expected to belong to the same class; otherwise, they are inconsistent. It is easy to verify whether the decisions are consistent or not by analyzing their decisions. However, it is unfeasible to compute equivalence classes with continuous features because the probability of samples with the same numerical value is very small. Intuitively speaking, the samples with the similar feature values should be classified into a single class in this case; otherwise, the decision is not consistent. According to this observation, neighborhood concept is introduced into the classical rough set theory and neighborhood rough set model was proposed to reduce attributes [18-19].

Definition 1 U is a nonempty and finite set of samples $\{x_1, x_2, ..., x_n\}$ (called a universe), A is a set of attributes $\{a_1, a_2, ..., a_m\}$. $< U, A >$ is called an information system. If $A = C \bigcup D$, where C is the set of condition attributes and D is the decision attributes, $<U,A=C\bigcup D>$ is called a decision system.

Definition 2 $<U,A=C\bigcup D>$ is a decision system, $\forall x_i\in U$, $B\subseteq C$, the neighborhood $\delta_B(x_i)$ of x_i in B is defined as: $\delta_B(x_i)=\{x_j\left|\Delta(x_j,x_i)\le\delta,x_j\in U\right.\}$ .where Δ is a distance function. $\forall x_1,x_2,x_3\in U$, Δ satisfies:

(1) $\Delta(x_1,x_2)\ge 0$, $\Delta(x_1,x_2)=0$ if and only if $x_1=x_2$;

(2) $\Delta(x_1,x_2)=\Delta(x_2,x_1)$;

(3) $\Delta(x_1,x_2)+\Delta(x_2,x_3)\ge\Delta(x_1,x_3)$.

Usually used distance functions include Manhattan distance, Euclidean distance and Chebychev distance. When x_1 and x_2 represent two samples in n-dimensional attribute space $C=\{c_1,\ldots,\mathrm{c}_n\}$, and $f(x,c_i)$ represents value of attribute c_i of sample x in ith dimension, Minkowsky distance can be defined as $\mathrm{D}_p(x_1,x_2)=(\overset{n}{\underset{i=1}{\text{å}}}|f(x_1,c_i)-f(x_2,c_i)|^p)^{1/p}$, here if $p=1$, call Manhattan distance D_1; if $p=2$, call Euclidean distance D_2; if $p=?$, call Chebychev distance $\mathrm{D}_{\text{¥}}$.

Definition 3 $U=\{x_1,x_2,\cdots,x_n\}$ is an universe, d is a real number and D is distance function, a neighborhood relation N on the universe U can be written as a relation matrix $U(N)=r_{ij}$, where $r_{ij}=\begin{cases}1,\Delta(x_i,x_j)\le\delta\\0,\Delta(x_i,x_j)>\delta\end{cases}$.

Definition 4 U is an universe and N is a neighborhood relation on U , $<U,N>$ is called a neighborhood approximation space. $\forall X\subseteq U$, lower and upper approximation of X in $<U,N>$ is defined as: $\underline{N}X=\{x_i\left|\delta(x_i)\subseteq X,x_i\in U\right.\}$, $\overline{N}X=\{x_i\left|\delta(x_i)\bigcap X\ne\varnothing,x_i\in U\right.\}$.

The boundary region of X in $<U,N>$ is defined as: $BN(X)=\overline{N}(X)-\underline{N}(X)$.

Definition 5 $<U,A=C\bigcup D>$ is a decision system and A can generate a neighborhood relation N on U , $<U,A=C\bigcup D,N>$ is called a neighborhood decision system.

Definition 6 $<U,A=C\bigcup D,N>$ is a neighborhood decision system. D divides U into N equivalence classes: $X_1,X_2,\ldots X_N$. $\forall B\subseteq C$, lower and upper approximation of decision D with respect to attributes B are defined as: $\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i$; $\overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i$.

where $\underline{N_B}X=\{x_i\left|\delta_B(x_i)\subseteq X,x_i\in U\right.\}$, $\overline{N_B}X=\{x_i\left|\delta_B(x_i)\bigcap X\ne\varnothing,x_i\in U\right.\}$.

The decision boundary region of D with respect to attributes B is defined as:

$BN(D)=\overline{N_B}(D)-\underline{N_B}(D)$.

The lower approximation of the decision is defined as the union of the lower approximation of each decision class. The lower approximation of the decision is also called the positive region of the decision, denoted by $POS_B(D)$. $POS_B(D)$ is the subset of objects whose neighborhood granules consistently belong to one of the decision classes.

Definition 7 $<U,A=C\bigcup D,N>$ is a neighborhood decision system, distance function Δ and neighborhood size δ , the dependency degree of D to B is defined as $\gamma_B(D)=\frac{|POS_B(D)|}{|U|}$ .where $|\cdot|$ is the cardinality of a ser. $\gamma_B(D)$ reflects the ability of B to approximate D .

Definition 8 $<U,A=C\bigcup D,N>$ is a neighborhood decision system, $B\subseteq C$,we say attribute B is a relative reduct if the following conditions are satisfied:

(1) Sufficient condition: $\gamma_B(D)=\gamma_A(D)$

(2) Necessary condition: $\forall a\in B,\gamma_B(D)>\gamma_{B-a}(D)$

The first condition guarantees that $POS_B(D) = POS_A(D)$ and the second condition shows there is not any superfluous attribute in a reduct. Therefore, a reduct is a minimal subset of attributes which has the same approximating power as the whole set of attributes.

As mentioned above, the dependency function reflects the approximating power of a condition attribute set. It can be used to measure the significance of a subset of attributes. The aim of attribute selection is to search a subset of attributes such that the classification problem has the maximal consistency in the selected feature spaces. In this section, we construct some measures for attribute evaluation, and then present greedy feature selection algorithms.

Definition 9 $< U, A = C \bigcup D, N >$ is a neighborhood decision system, $B \subseteq C$, $\forall a \in B$, one can defined the significance of a in B as $Sig_1(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D)$.

Note that the significance of an attribute is related with three variables: a , B and D . An attribute a may be of great significance in B_1 but of little significance in B_2 .What's more, the attribute's significance is different for each decision attribute if they are multiple decision attributes in a decision table. The above definition is applicable to backward feature selection, where redundant features are eliminated from the original set of features one by one. Similarly, a measure applicable to forward selection can be written as

$$Sig_2(a, B, D) = \gamma_{B \bigcup a}(D) - \gamma_B(D), \forall a \in A - B.$$

We say attribute a is superfluous in B with respect to D if $Sig_1(a, B, D) = 0$; otherwise, a is indispensable in B .

The objective of rough set based attribute reduction is to find a subset of attributes which has the same discriminating power as the original data and has not any redundant attribute. Although there usually are multiple reducts for a given decision table, in the most of applications, it is enough to find one of them. With the proposed measures, a forward greedy search algorithm for attribute reduction can be formulated as follows [18-19].

Algorithm1:Attribute reduction method based on neighborhood rough set model

```
Input: NDT =$< S, A = C \bigcup D, V, f >$ and neighbourhood $\delta$
Output: red;
(1)  red= $\varnothing$ ; // red is the pool to contain the selected attributes
(2)  for each $a_i \in C - red$
         computing $\gamma_{red \bigcup a_i}(D) = \frac{|POS_{B \bigcup a_i}(D)|}{|U|}$
         computing $Sig_2(a_i, red, D) = \gamma_{red \bigcup a_i}(D) - \gamma_{red}(D)$
     endfor
(3)  Selecting $a_k$ satisfying $Sig_2(a_k, red, D) = \max_i(Sig_2(a_i, red, D))$
(4)  if $Sig_2(a_k, red, D) > \varepsilon$ ,// $\varepsilon$ is a little positive real number use to control the convergence
       red=red$\bigcup a_k$
        go to (2)
     else
       return red
```

```
    endif
```

2.3 Teaching-Learning-Based Optimization

Teaching-Learning-Based optimization (TLBO) is a novel heuristic optimization algorithm base d on nature [21-23].The main idea of TLBO is to make use of the influence of a teacher on the output of learners in a class to achieve optimization purpose. The teacher is generally considered as a highly learned person who shares his or her knowledge with the learners. It is obvious that a good teacher trains learners such that they can have better results in terms of their marks or grades. The TLBO include two stages: teaching stage and learning stage. Teaching stage is that the learners (students) learn from teacher, and learning stage is that the learners (students) learn from one another.

GA, PSO and HS are most commonly optimization algorithm based on population and used widely in the field of optimization. However, algorithm parameters must be set in advance for these optimization algorithms. For example, the crossover probability, mutation rate and selection method are set in GA; Learning factors, the variation of weight and the maximum value of velocity must be set in PSO; the harmony memory consideration rate, pitch adjusting rate and number of improvisations must be set for HS. Many researches show algorithm parameters can usually affect highly optimization performance, but it is difficult that parameters are set correctly. Therefore, the widespread application of these optimization algorithms are limited. Comparison with above optimization algorithms, any algorithm parameters of need not to be set in TLBO. In addition, TLBO has the characteristics of simple principle, fast speed, high precision and better overall search ability.

In this paper, TLBO is applied to select a set of base classifiers from all the base classifiers to build an ensemble. The selection algorithm is given as follows.

Algorithm 2: Base classifiers selection based on TLBO

Input: Training set S *, Testing set* T *, all the base classifiers* $f_1, f_2, ..., f_D$ *and weight of base classifiers* $w_1, w_2, ... w_D$.

Output: base classifiers selected $f_{i_1}, f_{i_2}, ... f_{i_n} \in \{f_1, f_2, ..., f_D\}$*, and ensemble classification* $\hat{f}$.

Step 1: Initialize parameters.

population size NP ,number of generations G ,the number of all base classifiers D

Step 2 :Initialize the population

Using the formula $X = round(rand(1, D))$ *, we can randomly generate a population*

$$pop = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{NP} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & \vdots & \vdots & \vdots \\ x_{NP,1} & x_{NP,2} & \cdots & x_{NP,D} \end{bmatrix} .$$

where $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,D}\}$ *is a binary vector that represent the ith individual in pop,* $x_{i,j} \in \{0,1\}$ *. Each individual indicates a set of base classifiers selected. If the ith classifiers is selected ,the ith position of* X_i *is 1; while if the ith classifiers is not selected , the ith position of* X_i *is 0.*

Step 3: Calculate the fitness of each individual in pop .

According individual X_i *,a set of base classifiers are selected and ensembled by weighted voting, and the ensemble classification accuracy is expressed as* $f(X_i)$ *,that is fitness of the ith individual, so we calculate the*

```
fitness of all the individuals $fitness = \begin{matrix} f(X_1) \\ f(X_2) \\ \text{M} \\ f(X_{NP}) \end{matrix}$ .
Step 4: For i=1: G
    (1) "Teaching" phase
    (a) Calculate the difference.
        First ,the mean of population  pop is calculate and expressed as $M = [m_1, m_2, \text{K}\ m_D]$ ,where
            $m_i = \overset{NP}{\underset{k=1}{\text{å}}} x_{k,i} \Big/ NP$ ;
        Second ,find the best individual from pop as teacher $X_{teacher} = X_{i|f(X_i)=\max\{f(X_1), f(X_2), \text{K}\ f(X_{NP})\}}$ ;
        Third, the difference between $M$ and $X_{teacher}$ is expressed
                as $Difference = rand(1,D)\,?(X_{teacher} \quad TF\,?M)$ ,
            where $TF = round(1 + rand(1,D)(2-1))\ ?\ \{1,2\}$ ;
    (b) For  j=1: NP
            $X_j¢= X_j + Difference$ ;
            calculate fitness $f(X_j¢)$ ;
            if $f(X_j¢) > f(X_j)$
                $X_j = X_j¢$
            End if
        End For;
    (2) "Learning" phase
    For  j=1: NP
        Randomly select another individual $X_k$ ,such that $k\ ^1\ \ j$ ;
        If $f(X_j) > f(X_k)$
            $X_j^* = X_j + rand(1,D)\,?(X_j \quad X_k)$
        Else
            $X_j^* = X_j + rand(1,D)\,?(X_k \quad X_j)$
        End If
        Calculate fitness $f(X_j^*)$ ;
        if $f(X_j^*) > f(X_j)$
            $X_j = X_j^*$
        End if;
    End For;
  End For;
Step 5: Output base classifiers selected and ensemble classification.
```

A new population is generated after G times iteration, we find the best individual

$X_{best} = X_{i|f(X_i)=\max\{f(X_1),f(X_2),\text{K} f(X_{NP})\}}$ *and the best fitness*

$f(X_{best})$ *,where* $X_{best} = \{f_{i_1}, f_{i_2},..., f_{i_n}\} ? \{f_1, f_2,..., f_D\}$ *represent a set of base classifiers selected and*

$f(X_{best}) = \hat{f} = w_{i_1} f_{i_1} + w_{i_2} f_{i_2} + ... + w_{i_n} f_{i_n}$ *,(* $w_{i_1}, w_{i_2},..., w_{i_n}$ Î $\{w_1, w_2,..., w_D\}$ *)represent ensemble classification accuracy,.*

3. Our Proposed Method

Microarray data has the characteristics of small sample and high dimension, and contain a lot of irrelevant and redundant genes. Ensemble learning is an effective method for improving performance of classification. The diversity and accuracy are two important factors for affecting ensemble performance. How to increase diversity among base classifiers and accuracy of base classifiers is key problem for building an ensemble. In general, the diversity among base classifiers trained by using training set with higher diversity is more large, therefore producing training sets with high diversity is an effective method. Feature disturbance is effective to increase diversity among training sets, which different feature space with large diversity are produced by using feature disturbance method and then training subsets proposed have large diversity .In addition, in order to decrease noise to improve accuracy of base classifiers, irrelevant genes with classification task should be filtered.

This paper proposes a selective ensemble method to classify microarray data, and it includes four phases as follows:

(1) The first phase: in order to improve accuracy of classifier and decrease computation time, genes were reduced by using kruskal-wallis test and then training set is reduced to produce from original training set according to genes reduced.

(2) The second phase: multiple genes subsets with diversity are produced by using neighborhood rough set model with different radius, and corresponding training subsets are generated from training set reduced according to above genes subsets produced. Research shows the radius of neighborhood of NRS highly affect performance of NRS and different radius can obtain different reduction performance, therefore the diversity among training subsets obtained is more large.

(3) The third phase: above every training subset is used to train a classifier and then multiple base classifiers are generated. The base classifiers trained have high diversity because of diversity among training subsets.

(4) The fourth phase: a set of base classifiers are selected based on TLBO algorithm to build an ensemble classifier by weighted voting.

Figure 1 Exhibits Flow Chart of Our Proposed Method.

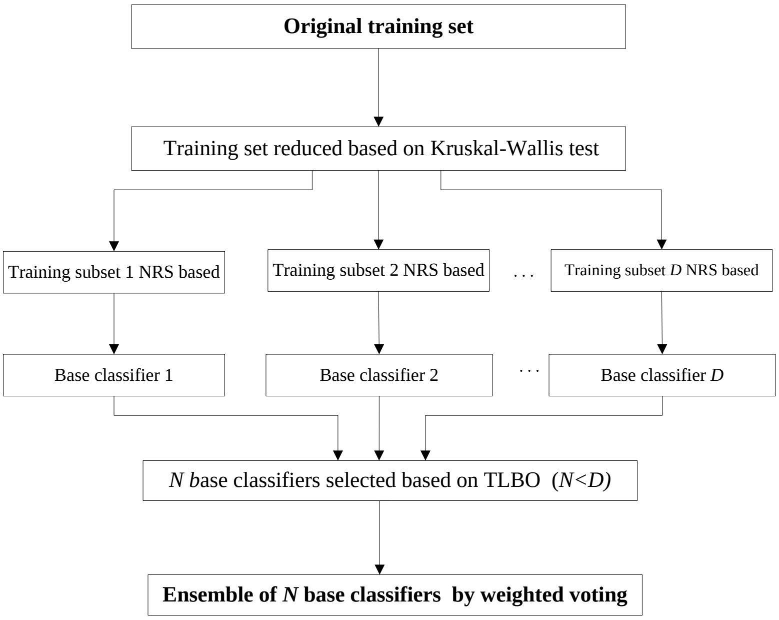

Original training set
Training set reduced based on Kruskal-Wallis test
Training subset 1 NRS based
Training subset 2 NRS based
. . .
Training subset D NRS based
Base classifier 1
Base classifier 2
. . .
Base classifier D
N base classifiers selected based on TLBO (N<D)
Ensemble of N base classifiers by weighted voting

Figure 1. Flow Chart of Our Proposed Method

4. Experiment

4.1 Experimental Datasets and Methods

In order to evaluate the performance of our proposed method, five well-known benchmarks tumor microarray datasets are selected and applied in our experiments. The characteristics of these datasets are shown in Table 1.

Table 1. Five Benchmark Tumor Microarray Datasets

DataSet	Classes	Genes	Samples	Training samples	Testing samples
CNS	2	7129	60	42	18
Leukemia	3	7129	72	38	34
Gliomas	2	12625	50	20	30
DLBCL	2	7129	77	32	45
ALL	6	12625	248	148	100

In order to explain effectiveness and superiority of our proposed method, five methods are selected and used for comparison with our method in our experiments.

Method 1: Original data (single classifier)
Method 2: Bagging
Method 3: AdaBoost
Method 4: Random Forest
Method 5: Kruskal-wallis+NRS
Method 6(Our proposed method): Kruskal-wallis+NRS+TLBO

To ensure the results of different methods does not happen by chance, the experiments are repeated 30 times independently, and results of 30 times are averaged as final results. In addition, RBF-SVM is employed as classifier in experiments.

4.2 Experimental Results and Analysis

For ensemble learning, the number of base classifiers usually affects performance of ensemble and it is difficult the number of base classifiers is determined correctly. In order to investigate the relationship between number of base classifiers and ensemble performance, the experiment are implemented when the number of base classifiers is equal to 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 in our experiments, respectively.

Table 2-6 displays the results of different methods when the number of base classifiers is 10, 20, 30, 40 and 50 respectively.

The final column of Table 2-6 gives the average number (*Num*) of base classifiers selected by using our proposed method. In addition, the "*best* "and "*average"* of method 6 are given in experiment because of randomness of TLBO , which represent the best results and average results of 30 times experiments, respectively. And standard deviation (*std*) is given to show stability of method 6.

Table 2. The Results of Different Methods (The Number of Base Classifiers D=10)

DataSet	Method1	Method2	Method3	Method4	Method5	Method6			
						best	*average*	*Std*	*Num*
CNS	66.67%	66.67%	50.00%	71.65%	66.67%	**77.78%**	**75.56%**	0.030	5
Leukemia	55.88%	64.71%	61.76%	73.28%	91.18%	**100%**	**98.24%**	0.016	5.4
Gliomas	66.67%	63.33%	66.67%	76.91%	80%	**90%**	**88%**	0.018	4.8
DLBCL	75.56%	84.44%	91.11%	84.24%	73.33%	**91.11%**	**87.56%**	0.046	4.2
ALL	68%	71%	76.00%	86.44%	94%	**96%**	**96.4%**	0.008	5
avg	66.56%	70.03%	69.11%	78.50%	81.03%	**90.98%**	**89.15%**	0.0236	4.88

Table 3. The Results of Different Methods (The Number of Base Classifiers D=20)

DataSet	Method1	Method2	Method3	Method4	Method5	Method6			
						best	*average*	*Std*	*Num*
CNS	66.67%	66.67%	66.67%	75.36%	61.11%	**83.33%**	**78.89%**	0.025	8.4
Leukemia	55.88%	70.59%	73.53%	80.1%	97.06%	**100%**	**100%**	0	8
Gliomas	66.67%	76.67%	70%	78%	70%	**93.33%**	**92%**	0.018	8.2
DLBCL	75.56%	88.89%	82.22%	86.27%	80%	**91.11%**	**89.78%**	0.03	6.2
ALL	68%	70%	80%	88%	89%	**98%**	**97.2%**	0.008	6.2
avg	66.56%	74.56%	74.48%	81.55%	79.43%	**93.15%**	**91.57%**	0.016	7.4

Table 4. The Results of Different Methods (The Number of Base Classifiers D=30)

DataSet	Method1	Method2	Method3	Method4	Method5	Method6			
						best	*average*	*Std*	*Num*
CNS	66.67%	66.67%	77.78%	76%	61.11%	**83.33%**	**81.11%**	0.03	10.6
Leukemia	55.88%	64.71%	79.41%	82.25%	91.18%	**100%**	**98.82%**	0.016	13.8
Gliomas	66.67%	70.00%	73.33%	71.24%	76.67%	**93.33%**	**91.33%**	0.018	11.8
DLBCL	75.56%	80.00%	82.22%	86.27%	75.56%	**91.11%**	**88.89%**	0.031	8
ALL	68%	71%	70.00%	88%	94%	**98%**	**97.2%**	0.004	10
avg	66.56%	70.48%	76.55%	80.75%	79.70%	**93.15%**	**91.47%**	0.0198	10.84

Table 5. The Results of Different Methods (The Number of Base Classifiers D=40)

DataSet	Method1	Method2	Method3	Method4	Method5	Method6			
						best	*average*	*Std*	*Num*
CNS	66.67%	66.67%	66.67%	76%	61.11%	**83.33%**	**80%**	0.03	12.2
Leukemia	55.88%	64.71%	61.76%	82.24%	97.06%	**100%**	**98.82%**	0.016	21.2
Gliomas	66.67%	76.67%	63.33%	71.24%	80%	**93.33%**	**91.33%**	0.018	13.8

DLBCL	75.56%	75.56%	64.44%	86.27%	80%	**91.11%**	**90.22%**	0.02	12
ALL	68%	72%	83.00%	88%	93%	**98%**	**97%**	0.007	11.4
avg	66.56%	71.12%	67.84%	80.75%	82.23%	**93.15%**	**91.47%**	0.0182	14.12

Table 6. The Results of Different Methods (The Number of Base Classifiers D=50)

DataSet	Method1	Method2	Method3	Method4	Method5	Method 6			
						best	*average*	*Std*	*Num*
CNS	66.67%	66.67%	55.56%	75.9%	66.67%	**83.33%**	**78.89%**	0.046	21.8
Leukemia	55.88%	64.71%	70.59%	82.36%	97.06%	**100%**	**99.41%**	0.013	20.8
Gliomas	66.67%	83.33%	66.67%	74%	80%	**93.33%**	**92%**	0.018	19.6
DLBCL	75.56%	80%	80%	80.01%	82.22%	**93.33%**	**91.56%**	0.01	14.6
ALL	68%	73%	82%	85%%	93%	**98%**	**97%**	0.01	17.4
avg	66.56%	73.54%	70.96%	79.45%%	83.79%	**93.59%**	**91.77%**	0.0194	18.84

From Table 2-6, we clearly see that our proposed method achieves the highest classification accuracy on all the datasets. The phenomenon reflected in Table 2-6 are very similar, and then conclusions are consistent.

The experimental results of Table 4 are analyzed as a representative when the number of base classifiers is 30. Table 4 shows the comparison of classification accuracy of different methods when the number of base classifiers is equal to 30.

In Table 4, it is obviously our proposed method achieves the highest classification accuracy on all the datasets and outperforms than other methods. For CNS, the accuracy achieved by method 6 (*"average"*) is 81.11%, which is 14.44% ,14.44%,3.33%,5.11% and 20% higher than that of method 1-4 and 5, respectively. For Leukemia, the accuracy achieved by method 6 (*"average"*) is 98.82%, which is 42.94% ,34.11%,19.41%,16.57% and 7.64% higher than that of method 1-4 and 5, respectively. For Gliomas, the accuracy achieved by method 6 (*"average"*) is 91.33%, which is 24.66%,21.33%,18%,20.09% and 14.66% higher than that of method 1-4 and 5, respectively. For DLBCL, the accuracy achieved by method 6 (*"average"*) is 88.89%, which is 13.33%,8.89%,6.67%,2.62% and 13.33% higher than that of method 1-4 and 5, respectively. For ALL, the accuracy achieved by method 6 (*"average"*) is 97.2%, which is 29.2%,26.2%,27.2%,9.2% and 3.2%,higher than that of method 1-4 and 5, respectively. The analysis indicates our proposed method is better than other methods, the reasons are as following: diversity among base classifiers trained by using NRS with different radius is large, and selective ensemble by using TLBO is effective for improving ensemble performance.

In addition, we obviously find that the accuracy of our proposed method outperform that of method 5 from Table 4. Comparison with method 5, the accuracy of method 6 (*"average"*) is improved 20%,7.64%,14.66%,13.33% and 3.2% on six datasets. The number of base classifiers selected by our proposed method from 30 base classifiers is about only 10.6,13.8,11.8,8 and 10, respectively. It indicates selective ensemble based on TLBO is effective to improve performance of ensemble algorithm, and can decrease memory costs and computation times.

In Table 4, "*avg*" represents summarized result which is calculates by averaging the accuracy over all datasets. The accuracy of our proposed method (*"average"*) is 91.47%, which is improved 24.91%,20.99%,14.92%,10.72% and 11.77% than method 1-4 and 5,respectively. The number of base classifiers selected by our proposed method is about only 36% (10.84/30). It indicates our proposed method is effective for classifying microarray data.

Figure 2 displays influence of number of base classifiers on classification accuracy by using our proposed method. We find the number of base classifiers highly affect classification accuracy and classification accuracy does not monotonously increase with the increase of number of base classifiers. The accuracy is worse when the number of base classifiers is 5, and then the accuracy quickly increases with the number of base classifiers, the accuracy basically stable when the number of base classifiers is about 20 to 40, finally accuracy slightly decreases when the number of base classifiers is about 45 to 50.It provides a reference to researchers for building an ensemble.

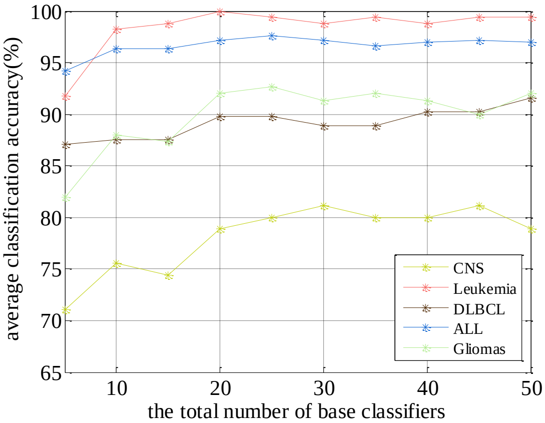


Figure 2. The Influence of Number of Base Classifiers on Classification Performance

Figure 3 shows the relation of the number of base classifiers selected by using our proposed method and total base classifiers. We can find that the number of base classifiers selected by our proposed method is only a few part of total base classifiers, and increases slightly with the increase of total base classifiers. Small number of base classifiers can improve computational speed and decrease storage requirements and it embodies the function of selective ensemble.

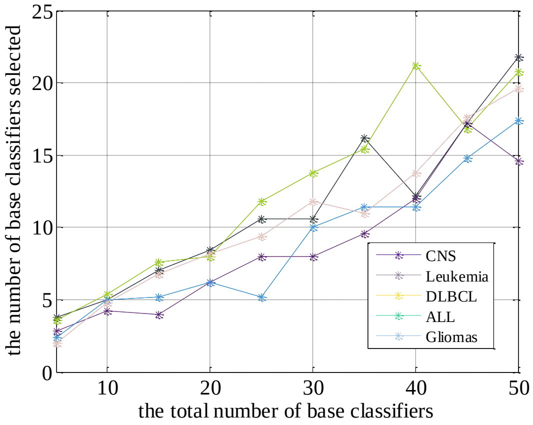


Figure 3. Variation of Number of Base Classifiers Selected and Total Base Classifiers

5. Conclusion

With the rapid development of high throughput technology, DNA microarray data are used to analyze gene levels in tumor cells. However, the imbalance of high dimension and small samples leads to limitation for analyzing microarray data. This paper proposes a selective ensemble method based on neighborhood rough set and teaching-learning-based optimization. Different feature subspaces are obtained by using neighborhood rough set with different radius on original training set, and training subsets with larger diversity are produced to build an ensemble on each feature subspace. TLBO is applied to select a set of base classifiers to build an ensemble, which can improve classification accuracy and decrease computation time and memory space. The experimental results indicate our proposed method is effective for microarray data classification.

Acknowledgements

This paper is supported by National Natural Science Foundation of China (NO: 11305097), Scientific Research Program Funded by Shaanxi Provincial Education Department (NO: 14JK1132) and Scientific Research Program Funded by Shaanxi University of Technology (NO: SLGKY14-03).

References

[1] M. B. Kursa, "Robustness of random forest-based gene selection methods," BMC bioinformatics, vol.15 no.1, **(2014)**, pp.1-8.

[2] J. Z. Wang, S. Zhou, Y. G. Yi and J. Kong, "An improved feature selection based on effective range for classification," The Scientific World Journal, **(2014)**.

[3] S. L. Wang, X. Li and S. Zhang, "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction," Computers in Biology and Medicine, vol.40 no.2, **(2010)**, pp.179-189.

[4] A. M. Bagirov, B. Ferguson, S. Ivkovic, G. Saunders and J. Yearwood, "New algorithm for multi-class tumor diagnosis using tumor gene expression signatures," Bioinformatics, vol.19, **(2003)**, pp.1800-1807.

[5] J. T. Horng, L. C. Wu and B. J. Liu, "An expert system to classify microarray gene expression data using gene selection by decision tree," Expert Systems with Applications, vol. 36 no.5, **(2009)**, pp. 9072-9081.

[6] P. Antal, G. Fannes and D. Timmerman, "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection," Artificial Intelligence in Medicine, vol. 29, **(2003)**, pp.39-60

[7] J. Ryu and S. B. Cho, "Gene expression classification using optimal feature classifier ensemble with negative correlation," Proceedings of the 2002 International Joint Conference on Neural Networks, Honolulu, **(2002)**, pp.198–203.

[8] N. Friedman, M. Linial and I. Nachman, "Using Bayesian networks to analyze expression data," Journal of computational biology, vol.7 no.3-4, **(2000)**, pp.601-620.

[9] L. Li, T. A. Darden and C. R. Weinberg, "Gene assessment and sample classification for gene expression data using a genetic algorithm k-nearest neighbor method," Combinatorial Chemistry & High Throughput Screening, no.4, **(2001)**, pp.727–739.

[10] D. Singh, P. G. Febbo and K. Ross, "Gene expression correlates of clinical prostate tumor behavior," Tumor Cell, no.1, **(2002)**, pp. 203–209.

[11] I. Guyon, J. Weston and S. M. D. Barnhill, "Gene selection for tumor classification using support vector machine," Machine Learning, vol. 46, **(2002)**, pp.389–422

[12] T. Chen, "Classification algorithm on gene expression profiles of tumor using neighborhood rough set and support vector machine," Advanced Materials Research, vol. 850-851, **(2014)**, pp.1238-1242

[13] H. Zhao, "Intrusion detection ensemble algorithm base d on bagging and neighborhood rough set," International Journal of Security and Its Applications, vol.7 no.5, **(2013)**, pp.193-204.

[14] T. Chen, Z. L. Hong, "A combined sum ensemble algorithm based on KICA and KFCM," Software Engineering and Knowledge Engineering: Theory and Practice, **(2012)**, pp.585-592.

[15] L. Shi, L. Xi and X. M., "A novel ensemble algorithm for biomedical classification based on ant colony optimization," Applied Soft Computing, vol.11 no.8, **(2011)**, pp.5674-5683.

[16] T. K. Ho, "The random subspace method for constructing decision forests," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.20 no.8, **(1998)**, pp. 832-844.

[17] R. D. Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," BMC bioinformatics, vol.7 no.1, **(2006)**, pp.1-13.

[18] Q. H. Hu, D. R. Yu and J. F. Liu, "Neighborhood rough set based heterogeneous feature subset selection," Information Sciences, vol.178 no.18, **(2008)**, pp. 3577-3594.

[19] Q. H. Hu, D. R. Yu and X. X. Zong, "Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation, "Journal of Software, vol.15 no.3, **(2008)**, pp.121-125.

[20] Z. H. Zhou and W. Tang, "Selective ensemble of decision trees," Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, **(2003)**, pp.476-483.

[21] R. V. Rao, V. J. Savsani and D. P. Vakharia, "Teaching-learning-based optimization: an optimization method for continuous non-linear large scale problems," Information Sciences, vol.183 no.1, **(2012)**, pp.1-15.

[22] R. V. Rao, V. J. Savsani and D. P. Vakharia, "Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems," Computer-Aided Design, vol.43 no.3, **(2011)**, pp. 303-315.

[23] T. Chen, "A selective ensemble classification method on microarray data," Journal of Chemical & Pharmaceutical Research, vol.6 no.6, **(2014)**, pp.2860-2866.

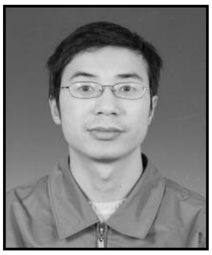
[24] J. Luts, A. Heerschap and J. A .K. Suykens, "A combined MRI and MRSI based multiclass system for brain tumor recognition using LS-SVMs with class probabilities and feature selection," Artificial Intelligence in Medicine, vol. 40 no.2, **(2007)**, pp.87-102.

Authors

Tao Chen, he received his Master from Lanzhou University (2007), and received his B.S. degree from Shaanxi University of Technology (2001), respectively. He is currently pursuing the Ph.D. degree in the School of Automation at Northwestern Polyechnical University. He is currently an associate professor in the school of mathematics and computer science, Shaanxi University of Technology. His current research interests are focused on data mining, pattern recognition and computational biology.

Zenglin Hong, he received his Ph.D. degree from School of Automation at Northwestern Polyechnical University. He is currently professor with School of Automation at Northwestern Polyechnical University. His current research interests are focused on systems engineering, land resources management and the regional economy.

Fang-an Deng, he received his Ph.D. degree from School of Science at Xidian University. He is currently professor of School of Mathematics and Computer Science at Shaanxi University of Technology. His current research interests are focused on intelligent information processing, rough set and soft algebra.



Xiao Yang, he received his B.S. degree from Hunan University (2002). He is currently an engineer in the school of mathematics and computer science, Shaanxi University of Technology. His current research interests are focused on data mining, pattern recognition and computational biology.

Jun Wei, he received his B.S. degree from Lanzhou University (2002). He is currently an experimenter in the school of mathematics and computer science, Shaanxi University of Technology. His current research interests are focused on data mining, pattern recognition and computational biology.

Man Cui, he received her Master in the school of communication engineering from Xi'an University of Science and Technology (2011), and received her B.E. degree from Xi'an University of Science and Technology (2008), respectively. She is currently pursuing the Ph.D. degree in the School of Automation at Northwestern Polytechnical University. Her current research interests are focused on decision support.