Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):2860-2866

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A selective ensemble classification method on microarray data

Tao Chen

School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong Shaanxi, China

---

ABSTRACT

For the characteristics of small samples and high dimension of microarray data, this paper proposes a selective ensemble method teaching-learning-based optimization based to classify microarray data. Firstly, in order to remove irrelevant genes with classification task, reliefF algorithm is used to reduce original gene set, and then a new training set is produced from orginal training set according to top-ranked genes obtained. Secondly, multiple bootstrap training subsets are produced based on bagging algorithm on above obtained training set to train base classifiers. Finally, multiple base classifiers are selected by using teaching-learning-based optimization to build an ensemble classifier. Experimental results on eight microarray datasets show our proposed method is effective and efficient for microarray data classification.

Key words: Microarray data; selective ensemble; reliefF; bagging; teaching-learning-based optimization

---

INTRODUCTION

The diagnosis of complex genetic diseases like cancer has conventionally been done based on the non-molecular characteristics like kind of tumor tissue, pathological characteristics and clinical phase. DNA microarray technology has concerned great attention in both the scientific and in industrial areas. Numerous examinations have been presented on the usage of microarray gene expression examination for molecular categorization of cancer. Several machine learning techniques have been used to classify microarray data [1].

However, due to the characteristics of small samples and high dimension of microarray data, and many existing irrelevant and redundant genes. It is leads to poor classification performance for most machine learning methods. In order to solve this problem and improve classification performance, ensemble technology was introduced to the area of data classification and obtain greatly success [2].

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem [3-5]. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use. An ensemble contains a number of learners which are usually called base learners. The generalization ability of an ensemble is usually much stronger than that of base learners. In 1995, Krogh indicated that the generalization error of ensemble is equal to average generalization error of individual minus the average differences of individual. Therefore, to enhance the generalization performance of ensemble, we should not only maximize the generalization ability of base classifiers, but also increase the differences between the various base classifiers [6-7]. Bagging [8] and boosting [9] are most common ensemble algorithms and achieve higher performance.

At present, most ensemble learning methods employ all base learners trained to build an ensemble. However, it leads to the increase of storage space and computation time, moreover the strategy of combining all base learners always does not achieve the best generalization performance. Selective ensemble is proposed to improve performance of ensemble method.

Teaching-learning-based optimization is a novel intelligent optimization algorithm based on population [10-11].TLBO simulates the behavior of teaching and learning in a class to improve academic performance of learners. Compare with genetic algorithm, particle swarm, harmony Search and differential Evolution, the biggest advantage of Tlbo is that it does not require any specific parameter to be set.Moreover, it has other several merits, such as fast convergent rate, simple principle and globe optimization, etc .

This paper proposes a selective ensemble method and it composes of three phases. The first phase is produces a new training set reduced from original training set by using reliefF algorithm [12]. In the second phase, multiple training subsets are produced by using bootstrap technology, and then multiple base classifiers are trained on above every training subset. The three phase, a set of base classifiers are selected by using teaching-learning-based optimization and combined to build an ensemble by weighted voting. In order to evaluate effectiveness of our proposed method, eight benchmark microarray datasets are selected and used in our experiment.

EXPERIMENTAL SECTION

ReliefF algorithm:

ReliefF is an extended and more robust version of the original Relief algorithm [12]. In contrast to many heuristic measures for feature selection, ReliefF does not assume conditional independence of the variables. The main idea of ReliefF is to estimate the quality of features based on how good their values discriminate between samples that are close. Consecutively random samples are drawn from the data set. Each time the k nearest neighbors of the same class and the opposite class are determined. Based on these neighboring cases the weights of the attributes are adjusted. As within the two previous algorithms the variables are ranked and different models are built by dropping the variable with the smallest weight. The remaining part of the selection procedure is completely analogous to the one followed in the two previous methods. Although the ReliefF algorithm is computationally more expensive and complex than the previous techniques, the cost of an exhaustive search is still much higher.

Bagging algorithm:

Bagging derived from bootstrap aggregation, was the first effective method of ensemble learning and is one of the simplest methods of arching [8]. The meta-algorithm, which is a special case of model averaging, was originally designed for classification and is usually applied to decision tree models, but it can be used with any type of model for classification or regression. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement. Each of these datasets is used to train a different model. The outputs of the models are combined by averaging (in the case of regression) or voting (in the case of classification) to create a single output.

Bagging trains a number of base learners each from a different bootstrap sample by calling a base learning algorithm. A bootstrap sample is obtained by subsampling the training data set with replacement, where the size of a sample is as the same as that of the training data set. Thus, for a bootstrap sample, some training examples may appear but some may not, where the probability that an example appears at least once is about 0.632. After obtaining the base learners, Bagging combines them by majority voting and the most-voted class is predicted.

Teaching-learning-based optimization:

Teaching-Learning-Based optimization (TLBO) is a novel heuristic optimization algorithm based on nature [10-11]. The main idea of TLBO is to make use of the effect of the influence of a teacher on the output of learners in a class to achieve optimization purpose. 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. The biggest advantage of Tlbo is that it does not require any specific parameter to be set, moreover, it has other several merits, such as fast convergent rate, simple principle and globe optimization, etc .

In this paper, a set of base classifiers are selected from all the base classifiers by using teaching-learning-based optimization and the algorithm is as follows.

Algorithm: 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 $f_{i_1}, f_{i_2}, ... f_{i_n}$ *, and ensemble classification*

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*

$pop = \begin{bmatrix} X_1 \\ X_2 \\ M \\ X_{NP} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & K & x_{1,D} \\ x_{2,1} & x_{2,2} & K & x_{2,D} \\ M & M & M & M \\ x_{NP,1} & x_{NP,2} & K & x_{NP,D} \end{bmatrix}$ *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* $fitness = \begin{bmatrix} f(X_1) \\ f(X_2) \\ M \\ f(X_{NP}) \end{bmatrix}$.

Step 4: For i=1: G

(1) "Teaching" phase

(a) Calculate the difference.

First ,the mean of population pop is calculate and expressed as $M = [m_1, m_2, \mathrm{K}\ m_D]$ *,where* $m_i = \sum_{k=1}^{NP} x_{k,i} \Big/ NP$ *;*

Second ,find the best individual from pop as teacher $X_{teacher} = X_{i|f(X_i)=\max\{f(X_1), f(X_2), \mathrm{K}\ f(X_{NP})\}}$ *;*

Third, the difference between M *and* $X_{teacher}$ *is expressed as* $Difference = rand(1,D) \times (X_{teacher} - TF \times M)$ *, where* $TF = round(1 + rand(1,D)(2-1)) \in \{1,2\}$ *;*

(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* $k \neq j$ *;*

If $f(X_j) > f(X_k)$

$X_j^* = X_j + rand(1,D) \times (X_j - X_k)$

Else

$X_j^* = X_j + rand(1,D) \times (X_k - X_j)$

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 $X_{best} = X_{i|f(X_i)=\max\{f(X_1), f(X_2), \mathrm{K}\ f(X_{NP})\}}$ *and the best fitness* $f(X_{best})$ *,where* X_{best} *represent a set of base classifiers selected and* $f(X_{best})$ *represent ensemble classification accuracy.*

Our proposed method:

Diversity among base classifiers and accuracy of base classifiers are key factors for affecting performance of

ensemble learning. The success of bagging algorithm is to produce training subsets with diversity by using bootstrap technology, therefore diversity of base classifiers is obtained by using bagging algorithm. For improving accuracy of base classifiers, ReliefF is effective method because it can remove irrelevant genes from original genes set to improve classification performance. In addition, TLBO is employed to select a set of base classifiers to build ensemble because of advantages of TLBO. According to above analysis, a selective ensemble based on TLBO is proposed for classifying microarray data. The concrete steps of our proposed method is given as follows.

Step1.Gene reduction

ReliefF algorithm is applied to remove irrelevant genes from original gene set, and a set of genes is build to produce a reduced training set from original training set. This step can improve accuracy of classification because of removing of irrelevant genes.

Step2. Production of base classifiers

Mutiple training subsets are obtained by using bootstrap technology to train classifiers. Because training subsets have much diversity, base classifiers trained have diversity.

Step3.Selection of base classifiers

A set of base classifiers are selected by using TLBO to decrease storage space and computation time.

Step4.Ensemble of base classifiers selected

Base classifiers selected are ensemble by weighted voting to classify new samples.

Experimental data and methods:

To evaluate performance of our proposed method, eight benchmark microarray datasets are selected and used in our experiments. The nine datasets are described in table 1.

Table 1 nine benchmark cancer microarray datasets

Data set	classes	genes	samples	training samples	testing samples
Colon	2	2000	62	43	19
Leukemia1	2	7129	72	38	34
DLBCL	2	7129	77	32	45
Gliomas	2	12625	50	20	30
Leukemia2	3	7129	72	38	34
MLLLeukemia	3	12582	72	27	45
SRBCT	4	2308	83	63	20
ALL	6	12625	248	148	100

In addition, in order to comparison superiority of our proposed method, four method (original, bagging, adaboost and ReliefF+bagging) are implemented. In our experiment, support vector machine with RBF (RBF-SVM) is employed as classifier. 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 experimental results.

Experimental results and analysis:

Many studies show the number of base classifiers in ensemble can also affect performance of ensemble algorithm. Therefore, the number of base classifiers is equal to10, 20, 30, 40 and 50 in our experiment, respectively.

Table 2-6 give the results of different methods on nine datasets when number of base classifiers is equal to 10,20,30,40 and 50, respectively. The “best” and “average” of our proposed method are given because randomness of TLBO. The “best” and “average” represents the best results and average results of 30 times experiments. The“std” represents standard deviation of 30 times experimental results. The “Num” represents average number of base classifiers selected by using TLBO in 30 times experiments.

We find that phenomenon of reflecting from table 2-6 is similar. It is easy to find that the classification accuracy of our proposed method is obviously higher than other methods. Especially, our proposed method outperforms ReliefF+bagging and it indicates selective ensemble based on TLBO is effective for improving performance of ensemble.

Table 4 is analyzed as a sample and results are given as follows. Table 4 displays the comparison of different methods on nine datasets when number of base classifiers is equal to 30.It is obviously that the classification accuracy of our proposed method is the highest in five methods according to table 4 and it indicates our proposed method is effective for microarray data classification.

For example, for colon, classification accuracy of our proposed method achieves 84.74%, which is improved at least

11.06% than other methods. For Leukemia1, classification accuracy of our proposed method achieves 85.59%, which is improved at least 3.24% than other methods. For DLBCL, classification accuracy of our proposed method achieves 96%, which is improved at least 9.33% than other methods. For Gliomas, classification accuracy of our proposed method achieves 86.33%, which is improved at least 9.66% than other methods. For Leukemia2, classification accuracy of our proposed method achieves 92.65%, which is improved at least 13.24% than other methods. For MLLLeukemia, classification accuracy of our proposed method achieves 92.67%, which is improved at least 8.23% than other methods. For ALL, classification accuracy of our proposed method achieves 92.44%, which is improved at least 7.79% than other methods. Only on SRBCT, classification accuracy of our proposed method and ReliefF+bagging are same and achieve 100%, which is improved at least 30% than other three methods. Compare with ReliefF+bagging method, the classification accuracy of our proposed method is improved 11.06%,9.12%,9.33%,9.66%,19.12%,8.23% and 3.6% on Colon, Leukemia1,DLBCL,Gliomas,Leukemia2, MLLLeukemia,ALL, respectively. For SRBCT, the result of the two methods are same. In general, our proposed method outperforms ReliefF+bagging and it indicates selective ensemble based on TLBO is effective for improving classification performance.

In table 4, "*avg*" represents summarized result which is calculates by averaging the accuracy over all datasets. The classification accuracy of our proposed method is the highest and achieves 91.82%, which is 27.2%, 22.94%,15.72% and 8.76% high than that of four methods, respectively. In addition, the average number of base classifiers selected for 30 is 9, about 0.3 (9/30).

Table 2 The results of different methods (the number of all the base classifiers is equal to 10)

DATASET	Original (%)	Bagging (%)	Adaboost (%)	ReliefF+ bagging(%)	Our proposed method			
					Best(%)	average(%)	std(%)	Num
Colon	63.16	63.1	78.95	73.68	89.47	82.63	3.55	4.2
Leukemia1	58.82	58.82	67.65	76.47	88.24	85.59	2.17	2.6
DLBCL	75.56	71.11	55.56	73.33	97.78	92.22	3.51	2.5
Gliomas	66.67	56.67	73.33	76.67	86.67	85.67	1.61	3.5
Leukemia2	55.88	61.76	64.71	73.53	94.12	87.65	4.96	2.4
MLLLeukemia	68.89	84.44	66.67	88.89	95.56	91.11	2.10	3.5
SRBCT	60	60.00	75.00	95.00	100.00	99.50	1.58	4.4
ALL	68	68	86.00	93.00	98.00	96.60	1.26	2.7
avg	64.62	65.50	70.98	81.32	93.73	90.12	2.59	3.2

Table 3 The results of different methods (the number of all the base classifiers is equal to 20)

DATASET	Original (%)	Bagging (%)	Adaboost (%)	ReliefF+ bagging(%)	Our proposed method			
					Best(%)	average(%)	std(%)	Num
Colon	63.16	63.16	73.68	73.68	89.47	83.68	2.99	8.7
Leukemia1	58.82	58.82	85.29	76.47	88.24	86.76	1.55	3.3
DLBCL	75.56	88.89	82.22	77.78	97.78	95.78	1.95	3.5
Gliomas	66.67	76.67	70.00	76.67	86.67	86.00	2.11	6.6
Leukemia2	55.88	70.59	73.53	85.29	97.06	90.88	4.03	4.9
MLLLeukemia	68.89	66.67	77.78	82.22	95.56	91.11	2.10	6.6
SRBCT	60	65.00	75.00	90.00	100.00	100.00	0.00	9.3
ALL	68	70	80.00	93.00	99.00	97.60	1.43	4.7
avg	64.62	69.98	77.19	81.89	94.22	91.48	2.02	6.0

Table 4 The results of different methods (the number of all the base classifiers is equal to 30)

DATASET	Original (%)	Bagging (%)	Adaboost (%)	ReliefF+ bagging(%)	Our proposed method			
					Best(%)	average(%)	std(%)	Num
Colon	63.16	63.16	73.68	73.68	89.47	84.74	4.61	11.1
Leukemia1	58.82	58.82	82.35	76.47	88.24	85.59	2.17	7.3
DLBCL	75.56	80.00	82.22	86.67	97.78	96.00	1.41	5.1
Gliomas	66.67	70.00	73.33	76.67	86.67	86.33	1.05	12.4
Leukemia2	55.88	64.71	79.41	73.53	97.06	92.65	4.85	6.2
MLLLeukemia	68.89	73.33	77.78	84.44	100.00	92.67	3.48	8.7
SRBCT	60	70.00	70.00	100.00	100.00	100.00	0.00	13.7
ALL	68	71	70.00	93.00	99.00	96.60	2.12	7.1
avg	64.62	68.88	76.10	83.06	94.78	91.82	2.46	9.0

Table 5 The results of different methods (the number of all the base classifiers is equal to 40)

DATASET	Original (%)	Bagging (%)	Adaboost (%)	ReliefF+ bagging(%)	Our proposed method			
					best(%)	average(%)	std(%)	Num
Colon	63.16	63.16	63.16	73.68	89.47	84.21	4.30	15.1
Leukemia1	58.82	58.82	64.71	76.47	85.29	83.24	1.42	15.8
DLBCL	75.56	75.56	64.44	80.00	95.56	92.00	4.22	10.4
Gliomas	66.67	76.67	63.33	73.33	86.67	86.33	1.05	14.7
Leukemia2	55.88	64.71	61.76	76.47	100.00	90.59	5.15	11.9
MLLLeukemia	68.89	73.33	71.11	88.89	91.11	90.00	1.17	18.2
SRBCT	60	65.00	80.00	95.00	100.00	100.00	0.00	21.3
ALL	68	72	83.00	93.00	99.00	97.20	2.15	8.9
avg	64.62	68.66	68.94	82.11	93.39	90.45	2.43	14.5

Table 6 The results of different methods (the number of all the base classifiers is equal to 50)

DATASET	Original (%)	Bagging (%)	Adaboost (%)	ReliefF+ bagging(%)	Our proposed method			
					best(%)	average(%)	std(%)	Num
Colon	63.16	63.16	78.95	73.68	89.47	84.21	2.48	21.1
Leukemia1	58.82	58.82	73.53	76.47	82.35	82.35	0.00	20.5
DLBCL	75.56	80.00	80.00	75.56	100.00	95.56	4.91	10.5
Gliomas	66.67	83.33	66.67	76.67	86.67	86.00	1.41	17.9
Leukemia2	55.88	64.71	70.59	73.53	97.06	89.12	4.17	15
MLLLeukemia	68.89	75.56	75.56	84.44	93.33	90.00	1.89	20.7
SRBCT	60	70.00	85.00	95.00	100.00	100.00	0.00	23.1
ALL	68	73	82.00	93.00	99.00	96.60	1.90	14.6
avg	64.62	71.07	76.54	81.04	93.49	90.48	2.09	17.9

Fig 1 displays the influence of number of base classifiers on classification accuracy. We find accuracy does not monotonously increase with number of base classifiers. The classification accuracy of our proposed method achieves the highest when number of base classifiers is about 20 or 30.

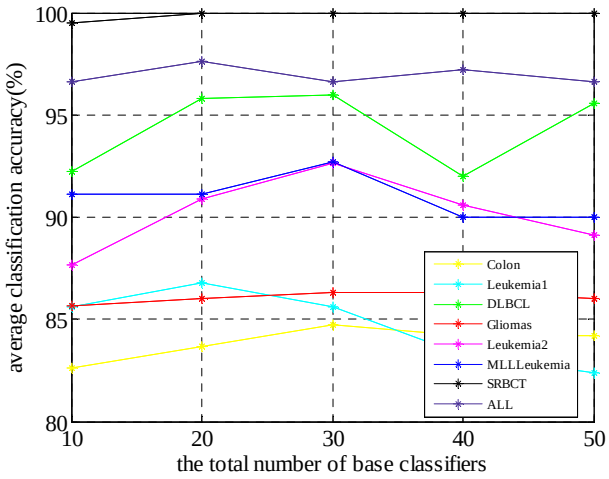

Fig 1 the influence of the number of base classifiers on classification performance

CONCLUSION

This paper proposes a selective ensemble method to classify microarray data. ReliefF algorithm is used to remove irrelevant genes to improve classification performance. Training subsets produced by bootstrap technology have large diversity and base classifiers trained have diversity. TLBO is applied to select a set of base classifiers to build an ensemble. Experimental results show our proposed method not only improve classification accuracy ,but also decrease computation time and storage space. Therefore our proposed method is effective and efficient for microarray data classification.

Acknowledgement
This paper is supported by National Natural Science Foundation of China (81160183, 11305097) and Scientific Research Program Funded by Shaanxi Provincial Education Department.

REFERENCES

[1] Wang Shu-Lin, Li X., Zhang S, et al. *Computers in Biology and Medicine*, **2010**, 40(2), 179-189
[2] Shi L, Xi L. , M X. ,et al, *Applied Soft Computing*, **2011**,11(8),5674-5683
[3] Zhao Hui. *International Journal of Security and Its Applications*, **2013**,7(5),193-204
[4] Chen Tao. *Application Research of Computers*.**2011**,28(1),139-141
[5] Zhou Zhi-Hua. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. **2003**, 476-483
[6] Chen Tao. *Journal of Computer Applications*,**2011**,31(5),1331-1335
[7] Chen Tao, Hong Zeng-Lin. *Software Engineering and Knowledge Engineering: Theory and Practice*, **2012**,585-592
[8] Breiman L. *Mach. Learn*. **1996**, 24(1),123–140
[9] Schapire R.. *Mach. Learn*. **1990** ,l5 (2),197–227
[10] Rao R V, Savsani V J, Vakharia D P. *Information Sciences*, **2012**,183(1),1-15
[11] Rao R V, Savsani V J, Vakharia D P. *Computer-Aided Design*, **2011**,43(3),303-315
[12]Kononenko I.*Proceedings of the European conference on machine learning, Lecture notes in computer science*.**1994**, 784,171-182