Bio-Medical Materials and Engineering 26 (2015) S1961–S1968
DOI 10.3233/BME-151499
IOS Press

A hybrid ensemble method based on double disturbance for classifying microarray data

Tao Chen[a,b,*], Huifeng Xue[a], Zenglin Hong[a], Man Cui[a] and Hui Zhao[b]
[a]*School of Automation, Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China*
[b]*School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong, 723000, Shaanxi, China*

Abstract. Microarray data has small samples and high dimension, and it contains a significant amount of irrelevant and redundant genes. This paper proposes a hybrid ensemble method based on double disturbance to improve classification performance. Firstly, original genes are ranked through reliefF algorithm and part of the genes are selected from the original genes set, and then a new training set is generated from the original training set according to the previously selected genes. Secondly, *D* bootstrap training subsets are produced from the previously generated training set by bootstrap technology. Thirdly, an attribute reduction method based on neighborhood mutual information with a different radius is used to reduce genes on each bootstrap training subset to produce new training subsets. Each new training subset is applied to train a base classifier. Finally, a part of the base classifiers are selected based on the teaching-learning-based optimization to build an ensemble by weighted voting. Experimental results on six benchmark cancer microarray datasets showed proposed method decreased ensemble size and obtained higher classification performance compared with Bagging, AdaBoost, and Random Forest.

Keywords: Microarray data, bagging, reliefF, neighborhood mutual information, teaching-learning-based optimization

1. Introduction

DNA microarray technology makes it possible to classify diseases according to gene expression levels in normal and tumor cells [1]. Ensemble learning is that multiple base classifiers are trained according to certain strategies, and then outputs of base classifiers are combined to classify new samples. Ensemble learning can improve classification performance by using other classifiers to average out the errors of another classifier. Therefore, ensemble learning can reduce the risk of selecting a poor performance classifier [2]. Krogh indicates that base classifier precision and diversity affect ensemble performance. Specifically, greater base classifier diversity can improve ensemble performance. Sample disturbance (Bagging [3] and Boosting [4]) and feature disturbance (Random Subspace [5] and Random Forest [6]) are effective methods for increasing base classifier diversity.

However, ensemble learning has three disadvantages. First, all the base classifiers need to be stored in ensemble and it may lead to require extra memory cost. Second, this process requires a significant amount of computation time for producing a new sample ensemble output during the prediction stage

* Address for correspondence: Tao Chen, School of Automation, Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, China. Tel.: +86 13109191838; Fax: +86 09162527378; E-mail: ct79hz@126.com.

[2]. Third, too many base classifiers may decrease base classifiers diversity, thereby reducing ensemble classification performance. Selective ensemble by selecting a set of base classifiers to build an ensemble can solve above problems. This process can improve classification performance and decrease memory costs and computation times.

Microarray data usually contains a large number of irrelevant and redundant genes, which decrease classification performance and increase algorithm complexity. Mutual information (MI) is widely used in gene selection [7]. MI measures the amount of information that one random variable contains about another random variable, which reflects the degree of linear or nonlinear dependency between variables. In the process of computing MI, probability distributions of variables and their joint distribution should been known. However, probability distributions are not usually known in practice. Neighborhood mutual information (NMI) is an effective measuring method to avoid MI disadvantages. It is constructed by integrating the concept of a neighborhood into Shannon's information theory, and it is a natural generalization of MI in numerical feature spaces [8, 9].

In 2011, R.V. Rao proposed a novel heuristic intelligent optimization algorithm nature based called teaching-learning-based optimization (TLBO).TLBO uses a teacher on the output of learners in a class to achieve optimization. Compared with GA and DE, TLBO doesn't require any parameters to be set in advance. In addition, it is simple, fast and provides better overall search ability [10].

This paper proposes a hybrid ensemble method. Firstly, a subset of genes is selected using a relief algorithm, and the original training set is reduced to produce a reduced training subset. Secondly, new training set is generated by sample disturbance based on bagging and feature disturbance based on neighborhood mutual information, and then multiple base classifiers are produced. Finally, a set of base classifiers are selected using TLBO to build an ensemble by weighted voting.

Section 2 presents the materials and methods, including to Relief algorithm, Neighborhood Mutual Information and Teaching-Learning-Based Optimization. Section 3 gives the basic ideas and steps of proposed method. Section 4 describes experiments on six benchmark microarray datasets and gives the experimental results and analysis. Section 5 holds the conclusion.

2. Materials and methods

2.1. ReliefF algorithm

ReliefF algorithm is proposed by Kononenko and it estimates the quality of attributes according to their values to distinguish between samples that are near each other. For that purpose, we selected a randomly samples x, relief algorithm searches for k nearest neighbors of x from the same class, called nearHist, and also k nearest neighbors of x from each of the different class, called nearMisses. The quality estimation $W(g)$ for each attribute g is update formula, nearHist and nearMisses. In the updated formula, the contributions of the hits and misses are averaged. The process is repeated n times to return the weights of all features. The relief algorithm was used in feature reduction because it is faster, fairly noise-tolerant, not limited by data types, and unaffected by feature interaction [11].

2.2. Feature selection algorithm based on neighborhood mutual information

Neighborhood mutual information is defined in the literature [8, 9]. $U = \{x_1, x_2, ..., x_n\}$ is a sample set and $F = \{f_1, f_2, \ldots f_m\}$ is a features set. $R, S \subseteq F$ are two feature subsets. The neighborhood of mutual

information of R and S is defined as $NMI_\delta(R;S) = -\frac{1}{n}\sum_{i=1}^{n}\log(\|\delta_R(x_i)\|\cdot\|\delta_S(x_i)\|/n\|\delta_{S\cup R}(x_i)\|)$, and $\delta(x) = \{x_i | \Delta(x, x_i) \le \delta\}$.

In order to obtain the features with higher NMI and deduce the impact of redundant features, a feature selection algorithm based on neighborhood mutual information and forward greedy search strategy was constructed as follows:

Input: samples set $U = \{x_1, x_2, ..., x_n\}$, features set $F = \{f_1, f_2, \ldots f_m\}$, decision attribute C, radius of neighborhood δ and the threshold of termination condition ξ.

Output: a reduct red.

Step 1: $red = \phi$;

Step 2: For each $f_i \in F - red$

(1) calculate $NMI_\delta(f_i \cup red; C)$; (2) calculate $Err(f_i, red, C) = NMI_\delta(f_i \cup red; C) - NMI_\delta(red; C)$;

Endfor

Step 3: Choose feature f_k which satisfies: $Err(f_k, red, C) = \max_i(Err(f_i, red, C))$;

Step 4: If $Err(f_k, red, C) > \xi$ (1) $red = red \cup f_k$; (2) goto Step 2; Else return and output red;

Endif

2.3. Teaching-learning-based optimization

Teaching-Learning-Based Optimization (TLBO), proposed by R.V. Rao in 2011, is a new heuristic optimization algorithm based on nature [10]. TLBO uses the effect of the influence of a teacher on the output of learners in a class to achieve optimization. The teacher is generally considered as a highly learned person who shares his or her knowledge with the learners. A good teacher trains learners so that they better understand the material and improve their marks or grades. The TLBO include two stages: teaching and learning. During the teaching stage, learners learn from the teacher; during the learning stage, learners learn from one another. GA and DE are the most common optimization algorithms. 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; the mutation operator and crossover operator must be set in DE. Research has shown that algorithm parameters can affect optimization performance. Correctly setting parameters is difficult, so the widespread application of these optimization algorithms is limited. TLBO does not require any preset algorithm parameters. In addition, TLBO is simple, fast, precise, and has better overall search ability.

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

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. We randomly generate a population $pop=[X_1,X_2,...,X_{NP}]'$ .where $X_i=\{x_{i,1},x_{i,2},...,x_{i,D}\}$ is a binary vector that represent the *i*th individual, $x_{i,j}\in\{0,1\}$. Each individual indicates a set of base classifiers selected. If the *i*th classifiers is selected, the *i*th position of X_i is 1; otherwise the *i*th 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 ensemble by weighted voting, and the ensemble classification accuracy is expressed as $f(X_i)$, that is the fitness of the *i*th individual, so we calculate the fitness of all the individual $fitness=[f(X_1),f(X_2),\ldots,f(X_{NP})]'$.

Step 4: For i=1: G

(1) Calculate the difference.

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

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

Third, the difference between M and $X_{teacher}$ is expressed as $Difference=rand(1,D)\cdot(X_{teacher}-TF\cdot M)$

(2) 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;

(3) For j=1: NP, Randomly select another individual X_k , such that $k\neq j$;

If $f(X_j)>f(X_k)$ $X_j^*=X_j+rand(1,D)\cdot(X_j-X_k)$, Else $X_j^*=X_j+rand(1,D)\cdot(X_k-X_j)$

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 iterations, the best individual $X_{best}=X_{i|f(X_i)=\max\{f(X_1),f(X_2),\ldots f(X_{NP})\}}$ and fitness $f(X_{best})$ are calculated, where X_{best} represents a set of base classifiers selected and $f(X_{best})$ represents ensemble classification accuracy.

3. Our proposed method

Ensemble learning is an effective method for improving classification performance. However, diversity and accuracy are two important factors for affecting ensemble performance. Increasing diversity and accuracy among base classifiers is a key problem for building an ensemble. In general, increasing training set diversity also increases base classifier diversity, thereby producing high diversity training sets. Bagging is a popular ensemble algorithm and has obtained great success in ensemble building. It generates training subsets from an original training set by using bootstrap technology to train base classifiers. Then, multiple base classifiers are combined to build an ensemble by majority voting. The bagging algorithm is successful because training subset diversity by bootstrap is increased [3]. In addition, feature disturbance also increases training set diversity.

Our method includes following four steps: (1) Features are reduced using reliefF, and the training set is obtained from the original training set. (2) Multiple bootstrap training subsets are produced by

Table 1

Six benchmark cancer microarray datasets

ID	Data set	classes	genes	samples	training samples	testing samples
Data1	Central Nervous system	2	7129	60	42	18
Data2	LeukemiaGloub	3	7129	72	38	34
Data3	MLLLeukemia	3	12582	72	27	45
Data4	Gliomas	2	12625	50	20	30
Data5	DLBCL	2	7129	77	32	45
Data6	ALL	6	12625	248	148	100

using a bagging algorithm from reduced training sets. (3) Features are selected by using NMI algorithm with different radius to produce a new training subset on each bootstrap training subset; multiple training subsets are produced, and then the base classifiers are trained on each produced training subset. Research has shown that the neighborhood radius affects NMI performance, and different radius can obtain different performances; therefore, NMI with different radius produces high diversity training subsets and base classifiers. (4) A set of base classifiers are selected by the TLBO algorithm to build an ensemble by weighted voting.

4. Experiment

4.1. Experimental datasets

Six well-known benchmark cancer microarray datasets are selected and implemented to evaluate the proposed method's effectiveness. Table 1 describes the characteristics of datasets.

4.2. Experimental methods and the parameter settings

Our method is compared against five other methods to highlight its effectiveness. The experiment is repeated 30 times independently, and the average results are used as the final results.

Method 1: SVM; Method 2: Bagging (decision trees); Method 3: AdaBoost (decision trees); Method 4: Random Forest (decision trees); Method 5: ReliefF + Bagging (SVM) + NMI (all the base classifiers are trained to ensemble; Method 6: The proposed method, relief + Bagging (SVM) + NMI + TLBO.

RBF-SVM is applied as a classifier in methods 1, 5 and 6. Methods 2, 3 and 4 apply decision trees as classifiers. For SVM, the gamma in the kernel function and the C of C-SVC are randomly selected. For random forest, the depth of each tree equals $\sqrt{n}$, where n is the number of features in the training set. In methods 5 and 6, the neighborhood radius in NMI is selected randomly from [0, 1]. TLBO in method 6 had a population size of 20 and 500 generations. In addition, to investigate the relationship between base classifier number and ensemble performance, the number of the base classifiers in ensemble equals 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 in experiments for method 2, 3, 4, 5 and 6.

4.3. Results and analysis

Due to space limitations, we do only list the results of the different methods when the number of base classifiers is 20 in Table 2. The final column of Table 2 gives the average number of base

Table 2

The results of different methods

Dataset ID	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6			
						best	*average*	*std*	*Num*
Data1	66.67%	66.67%	66.67%	75.36%	77.78%	88.89%	84.44%	0.024	7.6
Data2	55.88%	70.59%	73.53%	80.1%	88.24%	100%	98.24%	0.016	8.6
Data3	68.89%	66.67%	77.78%	83.24%	86.67%	100%	98.22%	0.018	8.6
Data4	66.67%	76.67%	70%	78%	63.33%	93.33%	89.33%	0.036	7
Data5	75.56%	88.89%	82.22%	86.27%	91.11%	100%	96.89%	0.025	7.6
Data6	68%	70%	80%	88%	90%	99%	98%	0.012	6.4
avg	66.95%	73.25%	75.03%	81.82%	82.86%	96.87%	94.19%	0.023	7.63

classifiers selected *(Num)* by using the proposed method. In addition, TLBO is a random optimization algorithm; therefore, the "best" and "average" of our proposed method are given, which represents the best result and average result of 30 times experiments, respectively. The standard deviation *(std)* is given to show the proposed method's stability. *"avg"* shows the summarized result, which is calculated by averaging the accuracy over all datasets.

From Table 2, we obtain the following results:

(1) Of the six methods, the proposed method has the highest classification accuracy for all the datasets. For the Central Nervous system, our proposed method produces the best average classification accuracy at 84.44%, while other methods produce averages of 66.67-77.78%. For LeukemiaGloub, the proposed method produces an average result of 98.24%, compared to 55.88-88.24% for other methods. For MLLLeukemia, the proposed method produces the best average result of 98.22%, compared to 66.67-86.67% for the others. For Gliomas, the proposed method produces the best average result of 89.33%, compared to 63.33-78% for the others. For DLBCL, the proposed method produces the best average result of 96.89%, compared to 75.56-91.11% for the others. For ALL, the proposed method produces the best average result of 98%, compared to 68-90% for the others.

(2) Compared with method 5, the proposed method obtains higher classification accuracy. The average accuracy of method 6 beats method 5 by about 6.66%, 10%, 11.55%, 26%, 5.78% and 8% across categories. Moreover, the number of base classifiers selected by our proposed method from 20 base classifiers is only about 7.6, 8.6, 8.6, 7.0, 7.6, and 6.4, respectively. This indicates that the selective ensemble based on TLBO improves ensemble classification accuracy.

(3) Method 5 outperforms method 4 on most datasets. The accuracy of method 5 is 2.42%, 8.14%, 3.43%, 4.84%, and 2% greater than that of method 4 on Central Nervous system, LeukemiaGloub, MLLLeukemia, DLBCL and ALL, respectively. On the whole, method 5 is better than method 4; it shows the hybrid ensemble method based on NMI and bagging is better than random forest. This is because base classifiers trained from different training subsets generated from different feature subsets via multi-radius NMI are more diverse.

(4) Method 5 is much more effective than methods 2 and 3. It indicates that feature disturbance based on NMI increases training subset diversity, and hybrid disturbance based on bagging and NMI is an effective method for building an ensemble.

Table 3 displays the average results when the number of base classifiers equals 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50, respectively. From Table 3, we see that our proposed method outperforms others methods on all the datasets. In addition, the average accuracy of our proposed method is 25.93%, 20.76%, 22.53%, 12.82%, and 10.26% greater than that of methods 1, 2, 3, 4, and 5, respectively. The

Table 3

The average results of different methods

Dataset ID	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6			
						best	*average*	*std*	*Num*
Data1	66.67%	66.67%	64.44%	74.98%	68.33%	85.56%	81.44%	0.034	12.08
Data2	55.88%	64.12%	68.53%	80.04%	91.76%	99.12%	97.35%	0.02	12.36
Data3	68.89%	75.56%	73.11%	82.62%	87.33%	98.89%	96.49%	0.024	12.02
Data4	66.67%	74.33%	66%	74.27%	68%	91.67%	87.2%	0.035	9.32
Data5	75.56%	81.33%	73.58%	84.61%	88.89%	99.33%	97.51%	0.018	10.32
Data6	68%	70.7%	76.44%	87.08%	91.4%	98.6%	97.28%	0.012	8.52
avg	66.95%	72.12%	70.35%	80.06%	82.62%	95.53%	92.88%	0.024	10.77

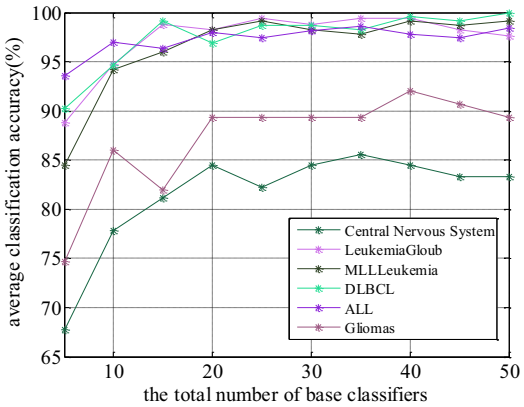

Fig. 1. The influence of the number of base classifiers on classification performance.

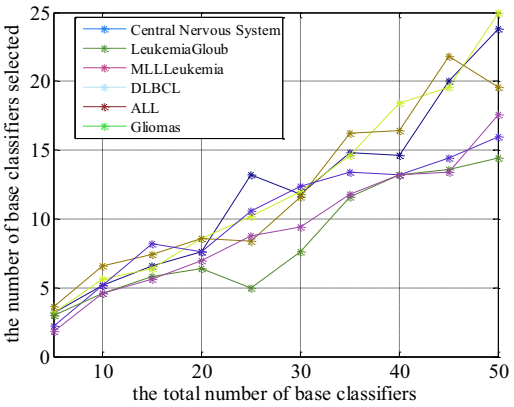

Fig. 2. Variation of number of base classifiers selected and total base classifiers.

number of selected base classifiers is much less than all the base classifiers, indicating that our method can achieve better classification accuracy by using a smaller number of base classifiers.

Figure 1 displays the influence of the number of base classifiers on classification accuracy in our proposed method. The base classifier number highly affects classification accuracy. The performance of our proposed method is worst when the number of base classifiers is 5; its performance quickly increases with the number of base classifiers, but its performance stabilizes at about 20-40 base classifiers. Its performance decreases gradually when the number of base classifiers is about 45-50.

Figure 2 displays the variation of the number of base classifiers selected with the total number of base classifiers by our proposed method. The number of base classifiers selected is only a few parts of all base classifiers, and increases slightly with the increase of all base classifiers. A small number of base classifiers can increase computational speed and decrease storage requirements.

5. Conclusion

Aim to the characteristic of microarray data, this paper proposes a hybrid ensemble method to improve classification performance. Bagging and NMI are used to enhance base classifier diversity. A set of base classifiers are selected to build an ensemble by using TLBO. The ensemble method improves classification performance, and decreases memory costs and computation times. The

experimental results indicate that our proposed method is effective for microarray data classification.

Acknowledgment

This paper is supported by National Natural Science Foundation of China (Nos: 11305097 and 11401357) and Scientific Research Program Funded by Shaanxi University of Technology (No: SLGKY13-44).

References

[1] M.B. Kursa, Robustness of random forest-based gene selection methods, BMC Bioinformatics **15** (2014), 1-8.
[2] L. Shi, L. Xi and X.M. Ma, A novel ensemble algorithm for biomedical classification based on ant colony optimization, Applied Soft Computing **11** (2011), 5674-5683.
[3] L. Breiman, Bagging predictors, Machine Learning **24** (1996), 123-140.
[4] R. Schapire, The strength of weak learn ability, Machine Learning **5** (1990), 197-227.
[5] T.K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998), 832-844.
[6] R. Díaz-Uriarte and S.A. De Andres, Gene selection and classification of microarray data using random forest, BMC Bioinformatics **7** (2006), 1-13.
[7] X. Liu, A. Krishnan and A. Mondry, An entropy-based gene selection method for cancer classification using microarray data, BMC Bioinformatics **6** (2005), 14-26.
[8] Q.H. Hu, W. Pan and S. An, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, International Journal of Machine Learning and Cybernetics **1** (2010), 63-74.
[9] Q.H. Hu, L. Zhang and D. Zhang, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Systems with Applications **38** (2011), 10737-10750.
[10] 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 **183** (2012), 1-15.
[11] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, Proceedings of the European Conference on Machine Learning, Lecture Notes in Computer Science **784** (1994), pp. 171-182.