Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):1597-1603

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Feature gene selection method based on improved harmony search algorithm

Jun Wei

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

ABSTRACT

DNA microarray data often contain tens of thousands of genes, where have a lot of irrelevant and redundant genes, these genes may seriously affect the accuracy and efficiency of classification. In order to solve this problem, this paper proposes a feature gene selection method based on improved harmony search algorithm. Firstly, genes are ranked using Relief F algorithm and preselected genes subset is obtained according to ranked-top genes .Then using improved harmony search algorithm to select feature genes from above genes subset. Finally we implement simulation experiments on three public microarray data sets. The results show that the proposed algorithm can achieve very high accuracy in the feature genes less, and is a effective and efficient feature genes selection algorithm. Especially, selected feature genes can help to understand microarray data.

Key words: microarray data; feature gene; Relief F algorithm; harmony search algorithm

INTRODUCTION

Gene chip [1] which also known as the DNA microarray is an advanced, large-scale, high-throughput detection technology , which has been widely and successfully applied in many fields of disease diagnosis, drug screening[2-3], and will be for human disease diagnosis, treatment and prevention of opening up new way, provide technical support platform for the rapid screening and pharmacogenomics in drug development studies of lead compounds.

Microarray data set contains tens of thousands of genes, but the number of samples are often less than one hundred. In the tens of thousands of genes, most of them have no obvious contribution for cancer classification, only a small amount of closely related genes have relationship with the classification task, and the noise and redundant genes will seriously affect the classification performance and efficiency. In order to figure out this problem, we choose the solution like using genes selection to eliminate the redundant and irrelevant genes, and will reduce the decrease the cost and improve the accuracy of clinical diagnosis, and this method also supply the reliable basis [4-5] for predicting disease .

At present, there are to two way to take feature gene selection which is filtration method(Filter) and winding method (Wrapper) [6-7]. The filtering method is usually adopted as a strategy to evaluate the relevance of each gene classification task, then order genes by the level of correlation and choose the higher ranking genes as the feature gene. The common filter method include "t- test" [8], "Fisher index" [9], "ReliefF"[10] and " classification index" [11]. It has many advantages like high efficiency, easy to implement, but it does not consider the interaction between each genes and will easily lead to produce the redundant genes. Wrapper method is usually adopted to evaluating classification algorithm for classification performance on a feature gene subset, then according to the evaluation result according to some strategy on set to adjust, in order to seek optimal factor set objective. Some heuristic search algorithm has been widely used in this field, such as genetic algorithm (Genetic Algorithm, [12-13] GA), particle swarm optimization (Particle Swarm Optimization, PSO) [14], ant colony algorithm (Ant Colony Optimization, ACO) [15]. Wrapper method has the advantages of good classification performance, less feature gene selection, but also have many disadvantages such as huge calculation, high time complexity, producing over fitting phenomenon in the high dimension, high noise data.

In order to solve this problem, this paper proposes a hybrid method of feature gene selection. The first stage is based on the ReliefF algorithm and calculate the correlation between each gene and categorical attributes. The second phase is using the improved harmony search algorithm to select feature gene. Voice search algorithm (Harmony Search, HS) is a new intelligent optimization algorithm, which simulates the process of the musicians to generate a wonderful harmony by repeatedly adjusting various musical tone. The algorithm has many advantages such as less adjustable parameters, a group of search capability, easy to merge with other algorithms. But the harmony search algorithm has some phenomenon with other intelligent algorithms which is premature. On the basis of [16] this paper, it will optimize harmony search algorithm, and will experiment on 3 public microarray datasets. The experiment results show that the algorithm is a feature gene selection algorithm wich have global search capability, high classification accuracy, and also could eliminate the noise and redundant gene. .

EXPERIMENTAL SECTION

Relief F 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.

Harmony algorithm:
The HS algorithm has been recently developed in an analogy with music improvisation process where musicians in an ensemble continue to polish their pitches in order to obtain better harmony. Jazz improvisation seeks to find musically pleasing harmony similar to the optimum design process which seeks to find optimum solution. The pitch of each musical instrument determines the aesthetic quality, just as the objective function value is determined by the set of values assigned to each decision variable. The steps in the procedure of classical harmony search algorithm are as follows:

Step 1. Initialize the problem and algorithm parameters. The optimization problem is specified as follows:

$$Minimize\ f(x) \qquad s.t. \qquad x_i \in X_i, i = 1, 2, \cdots, N$$

where $f(x)$ is an objective function; x is the set of each decision variable x_i ; N is the number of decision variables, X_i is the set of the possible range of values for each decision variable, $X_i : x_i^L \le X_i \le x_i^U$. The HS algorithm parameters are also specified in this step. These are the harmony memory size (HMS), or the number of solution vectors in the harmony memory; harmony memory considering rate (HMCR); pitch adjusting rate (PAR); and the number of improvisations(Tmax), or stopping criterion.

Step 2. Initialize the harmony memory. The HM matrix is filled with as many randomly generated solution vectors as the HMS

$$HM = \left[\begin{array}{c|c} x^1 & f(x^1) \\ x^2 & f(x^2) \\ \vdots & \vdots \\ x^{HMS} & f(x^{HMS}) \end{array}\right] = \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_N^1 & f(x^1) \\ x_1^2 & x_1^2 & \cdots & x_N^2 & f(x^2) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1^{HMS} & x_1^{HMS} & \cdots & x_N^{HMS} & f(x^{HMS}) \end{array}\right]$$

Step 3. Improvise a new harmony. Generating a new harmony is called ’improvisation’. A new harmony vector, $x^{'} = (x_1^{'}, x_2^{'}, \cdots, x_N^{'})$, is generated based on three rules: (1)memory consideration, (2)pitch adjustment ,(3)random selection. The procedure works as figure 1. $x_i^{'} = (x_1^{'}, x_2^{'}, \cdots, x_N^{'})$ is the *i*th component of x' , and $x_i^j\ (j = 1, 2, \cdots, HMS)$ is the *i*th component of the *j*th candidate solution vector in HM. Both r and $rand()$ are uniformly generated random number in the region of [0,1], and bw is an arbitrary distance bandwidth.

Step 4. Update harmony memory. If the new harmony vector, $x^{'} = (x_1^{'}, x_2^{'}, \cdots, x_N^{'})$ is better than the worst harmony in the HM, judged in terms of the objective function value, the new harmony is included in the HM and the existing worst harmony is excluded from the HM.

Step 5. Check stopping criterion. If the stopping criterion (maximum number of improvisations) is satisfied, computation is terminated. Otherwise, Steps 3 and 4 are repeated.

Our proposed method:
As with other intelligent algorithm, harmony search algorithm has premature phenomenon. In order to overcome the algorithm later stagnation, this article uses the literature in [16] algorithm, the introduction of the parameters, it increases with the number of iterations and reduce the worst at the early stage of the algorithm, and major updates and curry, late updates the best harmony harmony in the library, this kind of algorithm can effectively prevent the premature phenomenon, and can accelerate the the convergence process. The improved harmony search algorithm and specific steps are as follows:

Step 1: set up parameters: the number of variables N ; the maximum number of iterations $T_{\max}$; harmony memory size HMS ; tone tuning probability PAR ; the pitch adjusting bandwidth bw ; memory value probability $HMCR$.
Step 2: initialization of memory HM : According to $x_i = round(rand(1, N))$, the vector is randomly generated, which is composed of '0' and '1' and length is N .The '1' and '0' represent that the corresponding gene is selected or not selected.
Step 3: calculate each harmony fitness value in HM : training subset and testing subset are produced according x_i ,and then classification accuracy by using SVM on testing subset is as fitness value of x_i ,that is $f_i = f(x_i)$,and the best harmony x_{best} and worst harmony x_{worst} are found out.
Step 4: generate a random number $rand$ and compare with WSR . If $rand > WSR$,the worst harmony x_{worst} is selected, that is $x_{new} = x_{worst}$;otherwise the best harmony x_{best} is selected, that is $x_{new} = x_{best}$.
Step 5: generate a new variable. If $rand < HMCR$,the new variable comes from harmony memory; otherwise, if $rand < PAR$,the new variable is adjusted according to $x_{new} = round(x_{new} + (2 \times rand - 1) \times bw)$,
otherwise $x_{new} = round(rand(1, N))$.
Step 6: step 5 is repeated until all variables of new harmony are generated.
Step 7: update harmony memory. Firstly, the fitness of new harmony is calculated, that is $f_{new} = f(x_{new})$.
Secondly, if the worst harmony is selected, then f_{new} and f_{worst} are compared. If $f_{new} > f_{worst}$,then $x_{worst} = x_{new}$. if the best harmony is selected, then f_{new} and f_{best} are compared. If $f_{new} > f_{best}$,then $x_{best} = x_{new}$.
Step 8: check the algorithm termination conditions. If termination conditions is achieved, the best harmony is outputted, otherwise goto step 3.

EXPERIMENTAL DATA AND METHODS

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

Table 1 three benchmark cancer microarray datasets

Data set	classes	genes	samples	training samples	testing samples
DLBCL	2	7129	77	32	45
Leukemia	3	7129	72	38	34
ALL	6	12625	248	148	100

Experimental results and analysis:
In order to easy to study the RWBHS algorithm, it takes primary gene subset $N = 500$ and N=1000, and gives the result of freature gene selection method based on the classical harmony search algorithm (RHS). In order to avoid the influence of randomness of algorithm, RWBHS algorithm and RHS algorithm will runs 10 times.

(1) The classification accuracy
Figure 1 shows accuracy of the iterative process when algorithm in the $N = 500$ classification . We can observe the

convergence of RWBHS algorithm is superior than RHS algorithm from the chart. The average classification accuracy on 3 datasets are higher than RHS algorithm, it means that RWBHS algorithm has better global search capability.

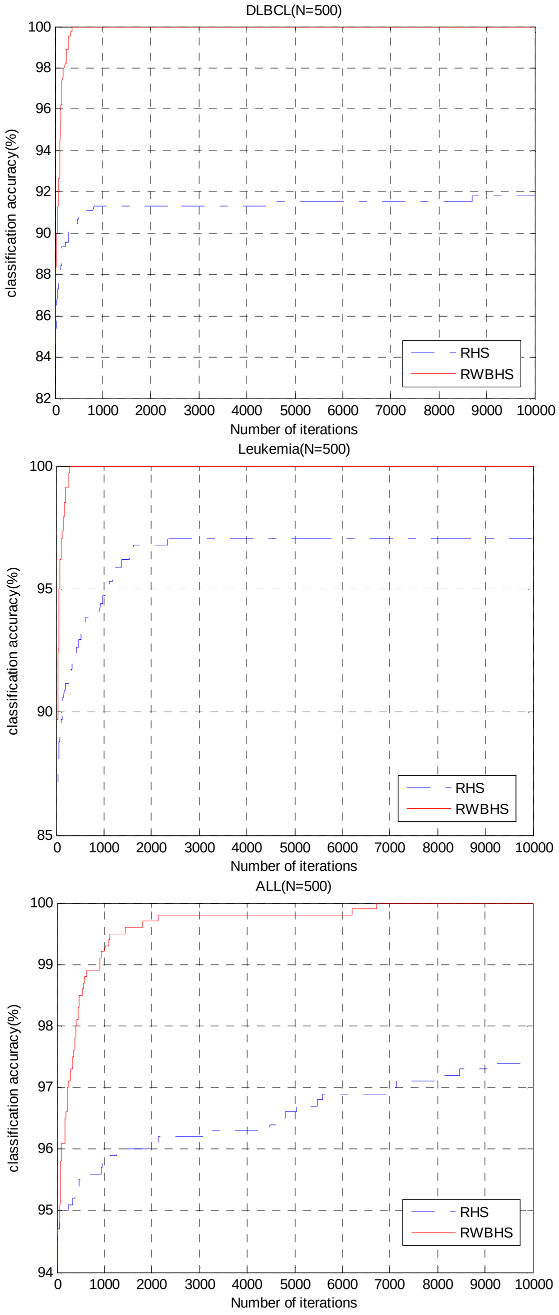

Figure 1 The average classification accuracy of iterative curve (N=500)

Table 2 Comparison of classification accuracy (N=500)

Data set	SVM(%)	RHS algorithm			RWBHS algorithm		
		Best(%)	Worst(%)	average(%)	Best(%)	Worst(%)	average(%)
DLBCL	75.6	93.3	91.1	91.8	100	100	100
Leukemia	55.9	97.1	94.1	96.8	100	100	100
ALL	68	98	97	97.4	100	100	100

Table 2 is the results of classification accuracy. We can observe that the accuracy rate of RWBHS algorithm can reach 100% in each experiment in the 3 data sets classification, it means that the algorithm has very strong stability.

(2) The number of feature gene

Figure 2 shows the iterative process feature gene subset algorithm of $N = 500$. We can see from the figure above, RWBHS algorithm convergence curve more smooth, the number of feature gene optimal feature search to the syndrome factor set was much less than that of RHS algorithm.

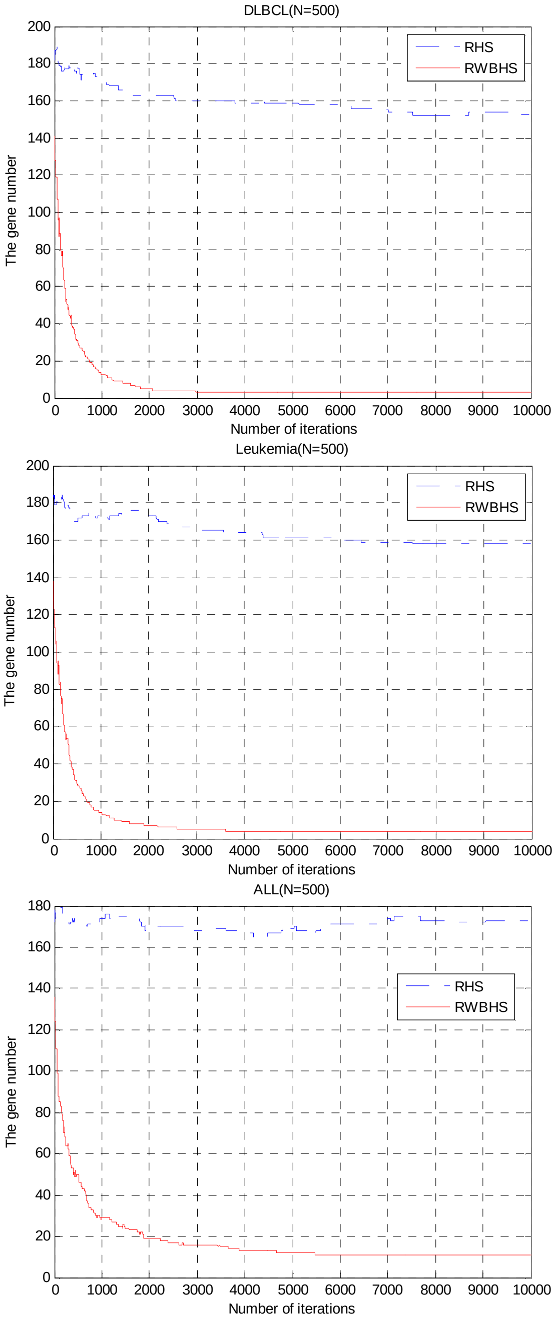

Figure 2 Average feature gene subset iterative curve (N=500)

Table 3 is the results of feature gene subset search to the. We can see, the RWBHS algorithm could effectively reject noise and irrelevant and redundant genes, and could search on the DLBCL data set of 2, on the Leukemia data set of 2, on the ALL data set of 7, please see table 4.

Table 3 feature gene subset numbers (N=500)

Data set	genes	RHS algorithm			RWBHS algorithm		
		Best	Worst	average	Best	Worst	average
DLBCL	7129	141	173	153	2	4	3
Leukemia	7129	152	167	158	2	5	4
ALL	12625	164	181	173	7	15	11

Table 4 the feature gene subset (N=500)

Data set	Num	Gene sequence (Series)
DLBCL	2	{555，1790}
Leukemia	2	{5543，1685}
ALL	7	{3331，6583,8615,8063,12305,6628,8556}

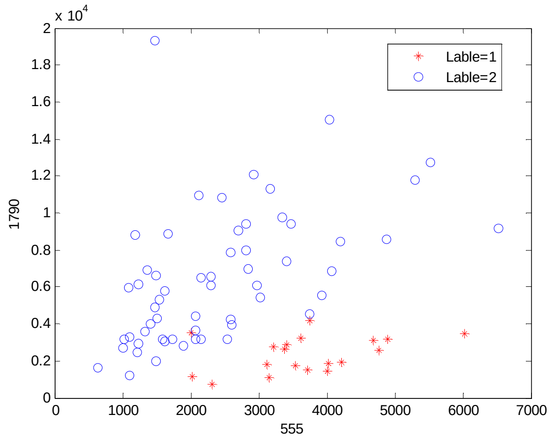

Figure 3 The classification results on DLBCL data sets

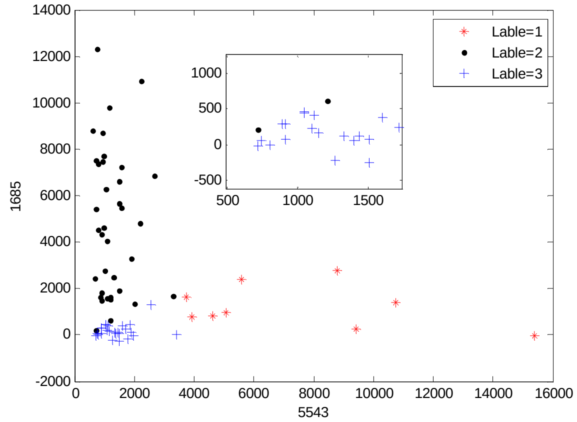

Figure 4 The classification results on Leukemia data sets

Figure 3 shows 2D {555,1790} scatter diagram which the optimal RWBHS algorithm searched in DLBCL data sets on the syndrome factor. Figure 4 shows 2D {5543,1685} scatter diagram which the optimal RWBHS algorithm searched in Leukemia data sets on the syndrome factor. The DLBCL data set which belongs to the 2 classification problems, Figure 3 shows that two sides of figure two class data clear and can be clearly separated. The Leukemia data set which belongs to the 3 classification problems, Figure 4 shows that first class data mainly concentrated in the lower right side of the graph, second kind of data are mainly concentrated in the left side of figure, and third kinds of data are mainly concentrated in the left side. By magnifying the diagram on the lower left side, it can be seen that the second class two data and third kinds of data are very close and easy to cause the classification error.

Based on the test results, we get the following conclusions: 1) RWBHS algorithm is better than RHS algorithm in classification accuracy. 2) The number of feature gene obtained by RWBHS algorithm are significantly less than RHS algorithm.

CONCLUSION

This paper presents a hybrid method of feature gene selection. The first stage is based on the ReliefF algorithm, the sort of microarray data set, the ranking of N genes constitute the primary gene subset, second phase using the improved harmony search algorithm to select feature gene. Through simulation experiments on 3 public microarray data sets, results show that the classification accuracy of the algorithm can reach 100%, the number of feature gene and search less, is a feature gene selection algorithm, worthy of further theoretical study.

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]Schena M,Shalon D,Davis R W,Brown P O. *Science*, **1995**, 270(5235): 467-470
[2]Ben-Dor A, Bruhn L, Friedman N, et al. *Journal of Computational Biology*, **2000**, 7(3-4): 559-583
[3]Wu Bin, Shen Ziyin. *Chinese Journal of Digest*, **2006**, 14(1): 68-74
[4]Tao Chen. *Journal of Chemical and Pharmaceutical Research*,**2014**,*6(6):15-28*
[5] Chen Tao. *Journal of Computer Applications*,**2011**,31(5),1331-1335
[6]Inza I,Larranaga P,Blanc R,et al. *Artificial Intelligence in Medicine*,**2004**,31(2):91-103
[7] Zhao Hui. *International Journal of Security and Its Applications*, **2013**,7(5),193-204
[8]Baldi P,Long A D.*Bioinformatics*,**2001**,17(16):509-519
[9] Furey T S,Cristianini N,Duffy N.*Bioinformatics*,**2000**,16(10):906-914
[10]Kononenko I. *Los Alamitos, CA: IEEE Computer Society*, **1994**: 171-182
[11] Rao R V, Savsani V J, Vakharia D P. *Computer-Aided Design*, **2011**,43(3),303-315
[12]PENG S H,XU Q H,FEN G X, et al. *FEB S Letters*, **2003**,555 (2) :358- 362 .
[13] Chen Tao. *Application Research of Computers*.**2011**,28(1),139-141
[14]SHE N Q , SHI W M , KON G W , e t a l . *Talanta*, **2007**, 71 (4) : 1679- 1683 .
[15]Chen Tao, Hong Zeng-Lin. *Software Engineering and Knowledge Engineering: Theory and Practice*, **2012**,585-592
[16] Longquan YONG. *Application of computer system*,**2011**,20(07) : 244-248.