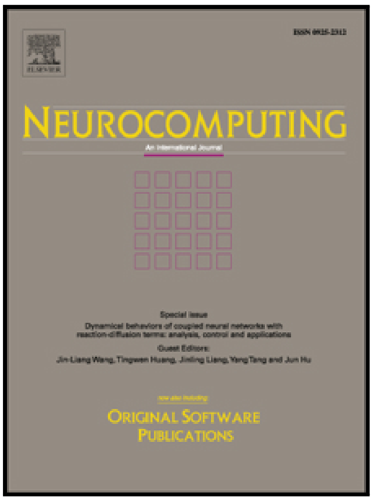

A Hybrid Feature Selection Algorithm for Gene Expression Data Classification

Huijuan Lu , Junying Chen , Ke Yan , Qun Jin , Yu Xue , Zhigang Gao

PII: S0925-2312(17)30415-0
DOI: 10.1016/j.neucom.2016.07.080
Reference: NEUCOM 18165

To appear in: *Neurocomputing*

Received date: 14 May 2016
Revised date: 16 June 2016
Accepted date: 16 July 2016

A Hybrid Feature Selection Algorithm for Gene Expression Data Classification

Huijuan Lu[a], Junying Chen[a], Ke Yan[a,*], Qun Jin[a,b], Yu Xue[c], Zhigang Gao[d]

[a] College of Information Engineering, China Jiliang University, 258 Xueyuan Street, Hangzhou, China, 310018.

[b] Department of Human Informatics and Cognitive Sciences, Waseda University, 2-579-15 Mikajima, Tokorozawa-shi, Saitama 359-1192, Japan.

[c] Nanjing University of Information Science & Technology, Nanjing, China, 210044.

[d] College of Computer Science，Hangzhou Dianzi University，Hangzhou, China, 310018.

Abstract

In the DNA microarray research field, the increasing sample size and feature dimension of the gene expression data prompt the development of an efficient and robust feature selection algorithm for gene expression data classification. In this study, we propose a hybrid feature selection algorithm that combines the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA). Experimental results show that the proposing MIMAGA-Selection method significantly reduces the dimension of gene expression data and removes the redundancies for classification. The reduced gene expression dataset provides highest classification accuracy compared to conventional feature selection algorithms. We also apply four different classifiers to the reduced dataset to demonstrate the robustness of the proposed MIMAGA-Selection algorithm.

[*] Corresponding author. Tel.: +86-153-9700-8303; fax: +86-571-8691-4580. E-mail address: yanke@cjlu.edu.cn

Keywords: Feature Selection; Mutual Information Maximization; Adaptive Genetic Algorithm; Gene Expression Data

1. Introduction

In bioinformatics, the DNA microarray technology is a benchmark technique for diagnosing cancers based on gene expression data [1, 2]. The clustering of the gene expression data provides a crucial way for identifying tumors [3, 4, 5]. However, the gene expression data is well-known as high-dimensional, large-scale and highly redundant data [6, 7]. Only a small number of genes are required in cancer diagnosis whereas the search space can be huge. Feature selection is therefore an important step to reduce both the dimension and redundancy of gene expression data during the classification process. An efficient and robust feature selection algorithm speeds up the learning process of classifiers and stabilizes the classification accuracy. In the gene expression data classification problem, two feature selection algorithms are commonly used, namely the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA).

Mutual information measures the correlation between two random data samples. In general, the mutual information describes the level of dependency between datasets. In statistics, all genes which belong to the same dataset are correlated. The most informative set of genes can be found by maximizing the mutual information of all genes belonging to the dataset [8].

Genetic algorithm (GA) is a parallel search heuristic, which is inspired by the natural selection process and the fundamental concepts in genetics [9]. Two operations are involved in the genetic algorithm, namely crossover and mutation, and corresponding to two probabilities: the crossover probability P_c and the mutation probability P_m. Inappropriate settings of P_c and P_m may result in various problems such as non-convergent or ‘premature convergence’ in search. The AGA improves the conventional GA by adjusting the values of P_c and P_m according to the search space variation. The adaptability of AGA makes it more robust and therefore enhances the likelihood of finding the global optimal solution.

Hybrid approaches combine two or more well-studied algorithms to form a new strategy to solve a particular problem. The hybrid approach usually capitalizes on the advantages from the sub-algorithms and therefore is more robust comparing with traditional approaches. Known hybrid approaches include ensemble classifiers [10, 11] and hybrid feature selection methods [12, 13].

In this study, we introduce a novel hybrid feature selection strategy combining the MIM and AGA to eliminate the redundant samples and reduce the dimension of the gene expression data. We demonstrate the effectiveness of the proposing MIMAGA feature selection method by comparing the classification accuracy rates with other existing feature selection methods. Then, four different classifiers are applied to the selected datasets to test the robustness of the proposing algorithm. All classifiers show classification accuracy rates higher than 80% (Section 4). We conclude the main contribution of our work as follows:

- **Novelty.** Both MIM and AGA are widely used feature selection algorithms in various fields. In bioinformatics, GA is more often used as the feature selection method in traditional gene classification problems. The hybrid approach that we introduced in this work has novel contribution to the literature.

- **Effectiveness.** The hybrid algorithm capitalizes on the advantages of the MIM and AGA. The genes selected by our algorithm show more accurate identification rates compared with existing feature selection approaches.
- **Robustness.** Four different classifiers are tested on the selected gene expression subsets. All classifiers produce stable classification accuracy curves in Section 4. And generally, the classification accuracy rates are all in acceptable region.

2. Related Work

The large-scale microarray gene expression technology provides a new way in cancer diagnosis [2]. By classifying the gene expression data, the top-most significant genes are discovered to provide useful information in cancer treatment. Feature selection is an important step to reduce the dimension and remove the redundancies of the gene expression data during the classification process. Tan et al. [14] introduced a feature selection framework which combines GA with various existing feature selection methods. They concluded that the hybrid methods are more effective and robust compared to each individual component algorithm. Ding and Peng [15] proposed a minimum redundancy maximum relevance (MRMR) feature selection framework to remove the redundancies in microarray gene expression data. Huang and Chow [16] introduced an effective feature selection scheme by estimating the mutual information based on a supervised data compression algorithm. Zhang et al. [17] employed the mutual information into multi-label classification problems and proved that the MIM effectively improved the classification accuracy of the multi-label classifiers. François et al. [18] improves the robustness of the forward feature selection by considering the mutual information criterion. Hoque et al. [19] proposed a greedy feature selection method using mutual information theory. The method combines both feature–feature mutual information and feature–class mutual information to find an optimal subset of features to minimize redundancy and to maximize relevance among features. In 2014, Wei et al. [20] integrated the MIM into a cloud computing system to perform classification for gene expression data. The program efficiency was greatly improved with almost the same classification accuracy.

In bioinformatics, data mining and machine learning, the GA is another effective feature selection algorithm that extracts the useful information from datasets; and multiple extensions of the conventional GA are proposed in the past decades [21, 22]. Ahmad et al. [23] introduced an improved hybrid genetic algorithm-multilayer perceptron network for intelligent medical disease diagnosis. Yun et al. [24] proposed a hybrid genetic algorithm approach for precedence-constrained sequencing problems. Silva et al. [25] used an extension of GA as a tool to identifying a subset of relevant genes and developing high-level classification rules for the cancer dataset NCI60, revealing concise and relevant information about the application domain. The accuracy of their methods was demonstrated to be higher than traditional approaches such as PART, J48, Naive Bayes, Random Forest and IBK. Diaz et al. [26] used GA to optimize the classification model in lung cancer diagnosis. Bagyamani et al. [27] introduced a hybrid GA for bi-clustering of gene expression data. Yang et al. [28] demonstrated the classification power of the Extreme learning machine (ELM) based on GA. The ELM is also utilized as the main classifier in this work.

The AGA algorithms extend GA by adjusting the crossover solutions and mutation variations. It becomes a more popular method applied to various fields. In 1994, Srinivas and Patnaik [29] first proposed to adjust both the crossover probability and mutation probability to get rid of the local minimum in search space. Hinterding et al. [30] introduced a self-adaptive genetic algorithm (SAGA) to iteratively search for the adapting level. Jakobović and Golub [31] demonstrated a 'self-contained' GA with steady-state selection. Qu et al. [32] proposed a co-evolutionary improved genetic algorithm (CIGA) for global path planning of multiple mobile robots. The improved GA algorithm adaptively avoids the local optimum problem and speeds up overall efficiency for searching. Chan et al. [33] applied AGA to the distributed production industrial area to eliminate the problem of search optimal crossover rates. The results showed that AGA largely improved the performance of genetic search. Kim et al. [34] combined AGA with fuzzy logic controller to solve a multiple scheduling problem. Wang and Tang [35] improved AGA based on hormone modulation mechanism and solved the job-shop scheduling problem by the improved method. Chen et al. [36] developed a forecasting algorithm based on support vector regression [37] and AGA.

3. A Hybrid Feature Selection Algorithm: MIMAGA-Selection

3.1 Mutual Information Maximization

Mutual information refers to the dependent information of one random sample (*x*) on the other random sample (*y*). For a given gene expression dataset, the overall mutual information can be expressed as:

$$I(X,Y)=\sum_{x\in S}\sum_{x\in T}p(x,y)\log_2\frac{p(x,y)}{p(x)\,p(y)}\ , \qquad (1)$$

where *p*(*x*) is the probability density of variable *x*, *p*(*y*) is the probability density of variable *y* and *p*(*x*,*y*) is the joint probability density. Suppose *N* represents the number of genes in the dataset, *A* represents the number of genes with gene expression profile *t* in class *c*, *B* represents the number of genes with gene expression profile *t* not in class *c*, *C* represents the number of genes without gene expression profile *t* in class *c*, *I*(*t*,*c*) represents mutual information of *t* of class *c*. Based on Formula (1), we have:

$$I(t,c)=\log\frac{p(t|c)}{p(t)}=\log\frac{p(t,c)}{p(t)\times p(c)}\approx\log\frac{A\times N}{(A+C)\times(A+B)} \quad . \qquad (2)$$

In Formula (2), if the gene expression profile *t* is irrelevant to class *c*, then *I*(*t*,*c*) = 0.

The maximum mutual information can be expressed as:

$$MaxMI(t)=\sum_{i=1}^{k}P(C_i|t)\log\frac{P(C_i|t)}{P(C_i)} \quad , \qquad (3)$$

where *k* represents the number of classes in the dataset.

The purpose of Mutual Information Maximization is to find genes that have strong dependency to all other genes in the same class. Applying MIM multiple times generally serves the purpose of genetic filtering.

3.2 Adaptive Genetic Algorithm

Crossover and mutation are two critical operations in GA. The crossover operation generates new individual in global. The mutation operation generates new individual in local. The two operations are mechanisms that endow GA with local and global search capability. The crossover probability (P_c) and the mutation probability (P_m) determine whether the GA algorithm converges to find the optimal solution. In the standard GA, P_c and P_m are pre-defined variables which are fixed in the GA search process. When P_c is too large, the global search is too coarse and the optimal solution can be missed out. When P_c is too small, the searching can be stuck in local minimal. When P_m is too large, the GA is similar to random search algorithms; and when P_m is smaller, the exploratory capability of the search is suppressed.

In order to find the most appropriate value for P_c and P_m, multiple cross-validations can be required. A more reasonable approach is to allow the GA adjusts the P_c and P_m during the searching process, which is called adaptive genetic algorithm (AGA). In AGA, the values of P_c and P_m can be adjusted following Formula (4) and (5):

$$P_c = \begin{cases} k_1 \dfrac{\left(f_{max} - f^{'} \right)}{\left(f_{max} - f_{avg} \right)} & , f^{'} \geq f_{avg} \\ k_2 \ , \ f^{'} < f_{avg} \end{cases} \tag{4}$$

$$P_m = \begin{cases} k_3 \dfrac{\left(f_{max} - f \right)}{\left(f_{max} - f_{avg} \right)} & , f \geq f_{avg} \\ k_4 \ , \ f < f_{avg} \end{cases} \tag{5}$$

In Equation (4), f_{max} represents the maximum of all individual fitness when AGA do a search operation, f_{avg} represents the average fitness, *f'* represents the bigger fitness of the parents in chromosome cross [38] and k_1, k_2, k_3, k_4 represent four control variables ranged from (0,1). The AGA optimization process is shown in Figure 1 and explained in detail in Section 3.3.

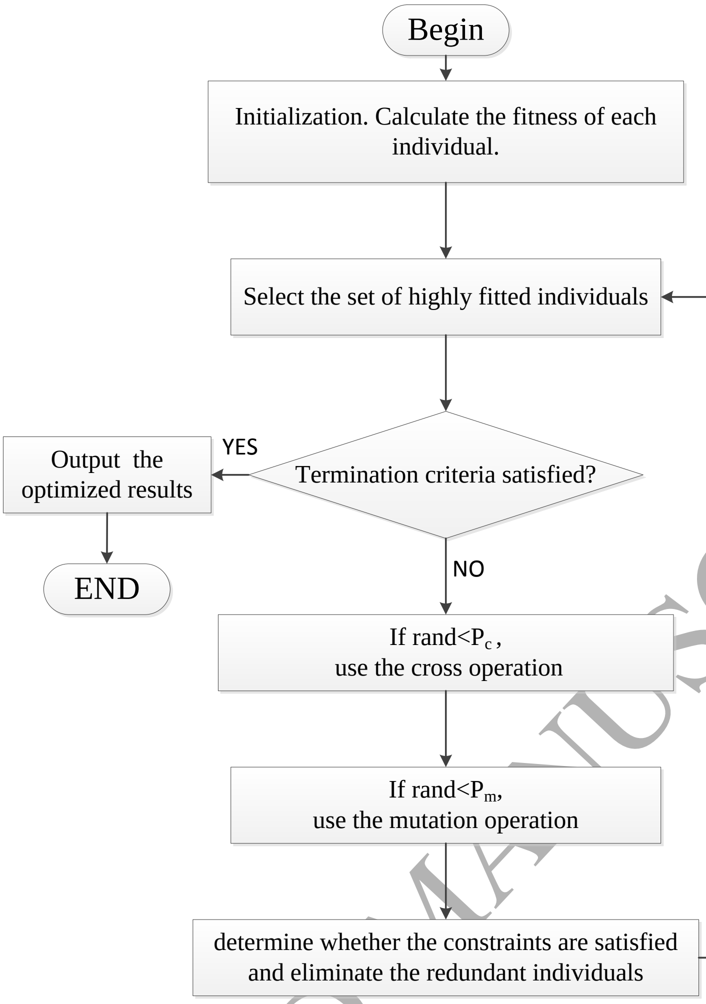
Begin
Initialization. Calculate the fitness of each individual.
Select the set of highly fitted individuals
Termination criteria satisfied?
YES
Output the optimized results
END
NO
If rand<P_c , use the cross operation
If rand<P_m, use the mutation operation
determine whether the constraints are satisfied and eliminate the redundant individuals

Figure 1. The AGA Optimization Process

3.3 MIMAGA-Selection

Combining MIM and AGA, we propose a gene selection algorithm named MIMAGA-Selection. By selecting the extreme leaning machine (ELM) as the classifier, the AGA Fitness becomes the ELM's classification accuracy. In Formula (4) and (5), we set k_1 = 0.9, k_2 = 0.6, k_3 = 0.1, k_4 = 0.001 and the number of maximal iteration loops to be 600. Suppose the gene expression dataset A has a_1 samples and a_2 genes. The detailed steps of MIMAGA-Selection algorithm can be described as follows:

(1) Calculate the mutual information of all genes in *A*. By applying MIM multiple times, we obtain a subset *B* of *A*. The gene number of *B* is set to be 300.

(2) Initialize the population for AGA and calculate the fitness for each individual. The population size is defined according to the problem space; the larger the size is, the easier the AGA searches for the optimal solution and the longer time will elapse. In this work, the population size *M* is set to 30. Each individual consists of several genes from *B*, and each gene has sample size a_1 .

(3) Adopt binary coding to encode 30 individuals in a population. After coding, each individual corresponds to a chromosome with length 300. (Chromosome is a row vector of size 300. If a slot takes the value from the *i*th column of *B*, the chromosome codes 1 to the *i*th bit. After the coding is completed for all slots of the active chromosome, the rest bits are set to 0).

(4) Calculate all fitness values for f_{max}, f_{avg}, f'.

(5) Select a set of highly fitted individuals by setting a threshold.

(6) Randomly paired the individuals in (5), according to the value of P_c in the Formula (4) using the crossover operation to generate new population.

(7) According to the value of P_m in the Formula (5), using the mutation operation to generate new population.

(8) Test whether the current optimal fitness value meets the target or the termination criteria are met. If yes, go to (9); otherwise, go to (4).

(9) Output the optimal subset of genes according to the decoding rules.

4. Experimental Results

Six gene expression datasets, namely Leukemia, Colon, Prostate, Lung, Breast and small-blue-round-cell tumor (SBRCT) are tested in this experiment. The sample number, gene number and labeled classes are summarized in Table 1. Among all datasets, only the SRBCT dataset is a four-class dataset, the rest datasets are binary.

Datasets	Sample Num	Gene Num	Distribution	
			Class	Sample
Leukemia	34	7130	ALL	20
			AML	14
Colon	62	2000	Negative	31
			Positive	31
Prostate	34	12600	Negtive	25
			Positive	9
Lung	149	12535	Negtive	134
			Positive	15
Breast	19	24482	Non-relapse	7
			relapse	12
SRBCT	63	2309	EWS	23
			RMS	20
			NB	12
			NHL	8

Table 1. Gene expression datasets

For each of the six gene expression datasets, we perform MIMAGA-Selection nine times with different target number of selected genes. The results are shown in Table 2.

Datasets	Number of Genes								
	1	2	3	4	5	6	7	8	9
Colon	19	57	77	107	136	149	171	187	202
Leukemia	7	44	60	91	125	148	164	177	198
Prostate	3	34	60	93	118	153	166	186	205
Lung	3	42	74	89	122	151	170	186	216
Breast	6	23	59	80	125	140	158	168	216
SRBCT	28	30	78	97	115	145	169	194	207

Table 2. The number of genes after applying MIM-AGA Selection to the seven gene expression datasets

The classification accuracy rates for each subsets using ELM are shown in Table 3. It is noted that each classification accuracy rate is an average result in repeating 30 times of the classification process.

Datasets	Classification accuracy rates %								
	1	2	3	4	5	6	7	8	9
Leukemia	97.62	96.67	95.95	96.96	97.14	95.95	94.09	97.14	97.14
Colon	89.09	81.82	85.45	80.4	81.82	81.82	83.18	86.90	83.41
Prostate	96.54	97.12	97.12	97.69	97.31	96.54	96.73	97.12	97.31
Lung	97.80	92.00	93.57	92.78	94.43	94.89	93.22	93.33	94.67
Breast	82.47	84.32	87.19	85.12	84.39	86.73	92.31	94.37	95.21
SRBCT	94.66	95.80	90.11	89.09	86.36	87.16	88.07	88.98	88.64

Table 3. The classification accuracy rates by MIMAGA-Selection and ELM

To demonstrate the effectiveness of the MIMAGA-Selection algorithm, we apply three existing feature selection algorithms: ReliefF [39, 40], sequential forward selection (SFS) [41, 42] and MIM on the same datasets with the same target gene numbers. The ELM with the same setting is applied to the selected gene subsets of the three algorithms. The classification accuracy rates are shown in Tables 4-6.

Datasets	Classification accuracy rates %								
	1	2	3	4	5	6	7	8	9
Leukemia	62.50	64.55	65.45	68.18	70.42	66.25	63.75	61.67	60.42
Colon	64.55	66.82	68.18	60.83	62.08	65.42	68.33	63.75	67.08
Prostate	55.91	57.50	59.17	60.42	61.15	59.62	58.85	53.46	54.62
Lung	50.54	51.54	53.08	54.23	5.925	58.57	57.50	54.29	50.71
Breast	50.71	51.67	52.33	54.33	53.44	52.81	51.25	50.94	50.31
SRBCT	58.33	59.17	68.03	62.50	65.38	64.23	63.46	60.38	59.62

Table 4. The classification accuracy rates by ReliefF and ELM

Datasets	Classification accuracy rates %								
	1	2	3	4	5	6	7	8	9
Leukemia	96.88	95.45	93.64	90.53	87.43	85.34	93.60	94.54	95.76
Colon	52.11	63.10	65.17	64.21	64.28	63.18	61.38	67.78	70.63
Prostate	83.98	82.94	81.63	83.28	84.12	82.21	83.29	84.28	86.28
Lung	83.27	84.21	81.77	83.27	86.90	87.27	82.38	84.29	89.57
Breast	70.22	73.58	74.48	76.38	77.28	78.59	78.94	70.29	74.22
SRBCT	81.47	86.78	85.29	86.66	82.07	79.26	80.27	83.42	80.32

Table 5. The classification accuracy rates by SFS and ELM

Datasets	Classification accuracy rates %								
	1	2	3	4	5	6	7	8	9
Leukemia	72.67	68.83	76.83	68.67	74.50	76.50	69.33	69.00	70.83
Colon	65.33	66.50	63.50	73.83	62.33	65.45	66.33	63.33	68.17
Prostate	86.50	85.00	86.83	84.17	85.17	84.83	88.67	84.50	83.83
Lung	79.52	77.94	77.22	77.14	78.33	77.22	78.50	77.61	77.62
Breast	80.00	70.59	73.56	72.31	75.65	73.21	76.33	73.89	73.43
SRBCT	86.82	87.30	77.78	79.37	85.71	80.95	79.36	79.68	78.73

Table 6. The classification accuracy rates by MIM and ELM

In general, the genes selected by the MIMAGA-Selection algorithm provide higher classification accuracy rates compared to existing feature selection algorithm. We demonstrate the classification accuracy comparisons for Leukemia, Colon and SRBCT datasets in Figure 2, 3 and 4 respectively.

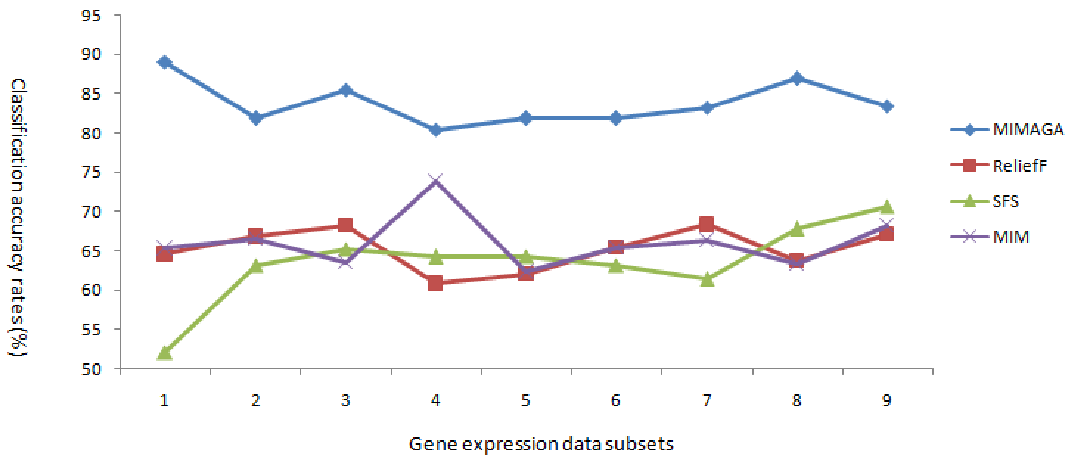

Figure 2. Classification accuracy rates using different feature selection algorithms on the Leukemia dataset

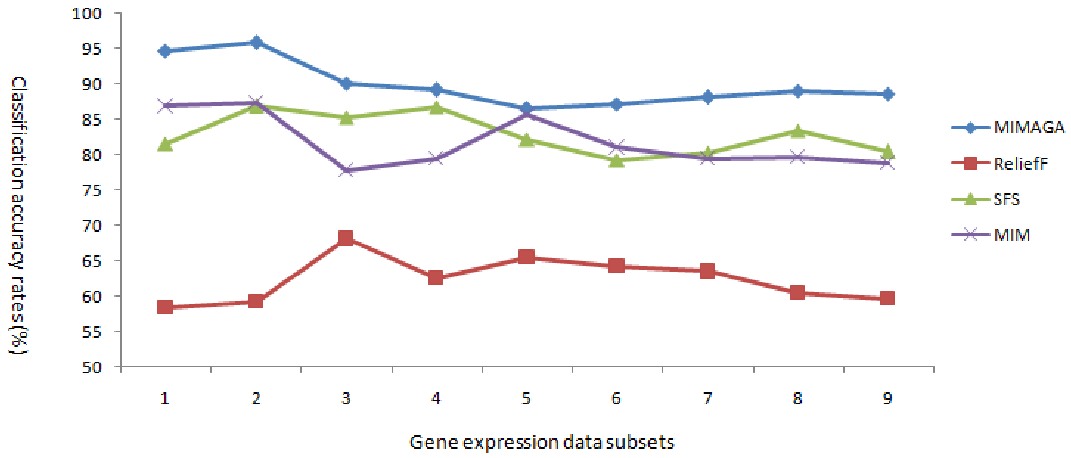

Figure 3. Classification accuracy rates using different feature selection algorithms on the Colon dataset

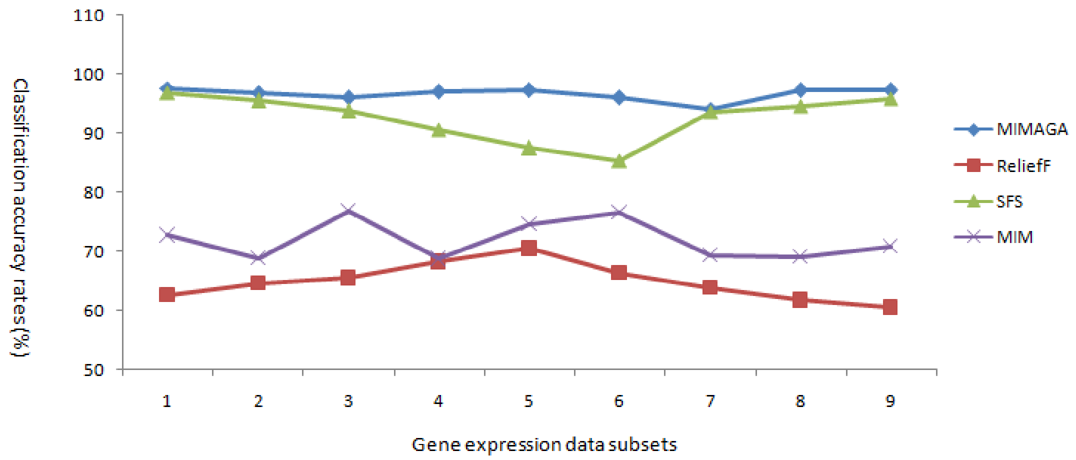

Figure 4. Classification accuracy rates using different feature selection algorithms on the SRBCT dataset

To further demonstrate the effectiveness of the selected genes from the MIMAGA-Selection algorithm, we classify the MIMAGA-Selection selected gene expression data subsets using four different classifiers, namely the back propagation neural network (BP), support vector machine (SVM), ELM and regularized extreme learning machine (RELM) [43]. In BP, the level of structure, the number of nodes in the hidden layer, the maximum iteration loops are set to be 3, 50, 600 respectively; and the kernel function is Sigmoid. For SVM, the penalty coefficient and the gamma value are 0.12 and 0.13; and the kernel function is Sigmoid. The settings for ELM and RELM are the same. The classification accuracies for Prostrate, Lung and Breast datasets are shown in Figure 5, 6 and 7 respectively.

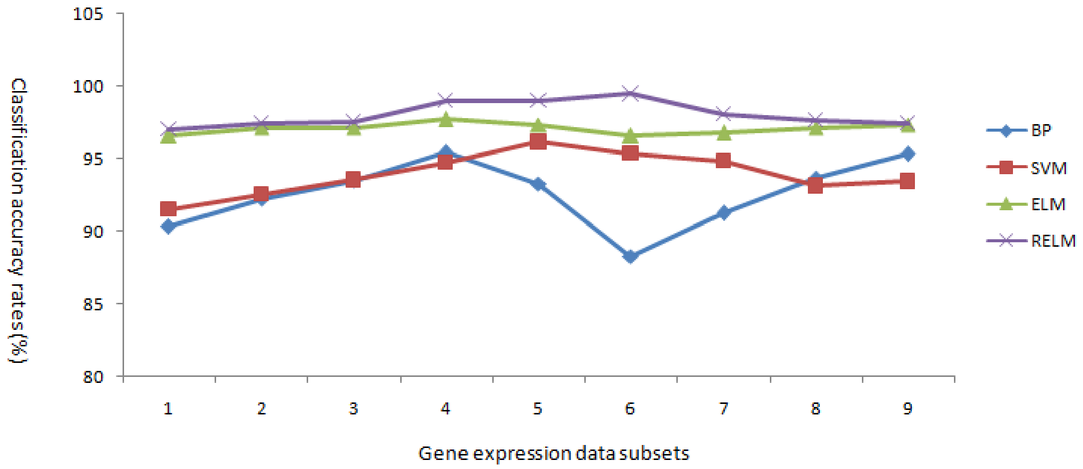

Figure 5. Classification accuracy rates using different classifiers on the Prostrate dataset

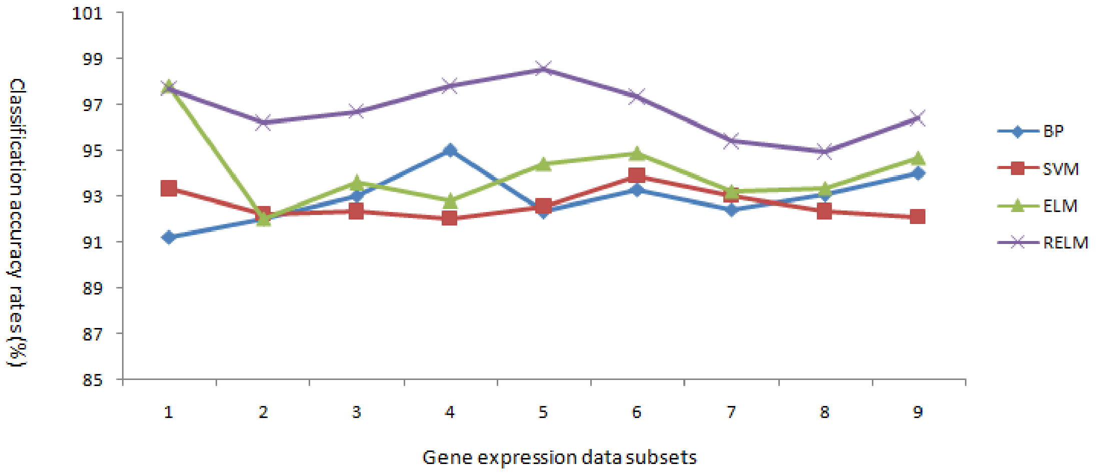

Figure 6. Classification accuracy rates using different classifiers on the Lung dataset

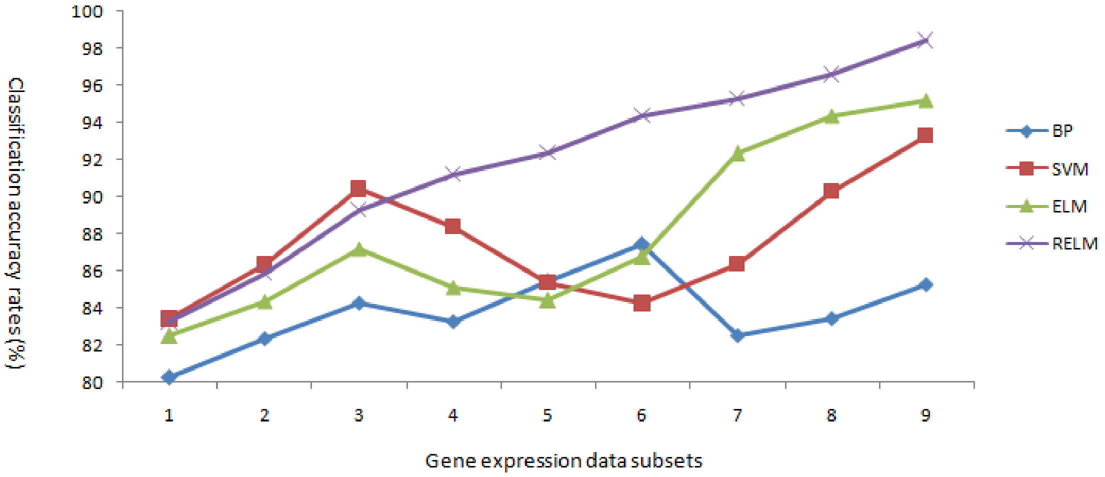

Figure 7. Classification accuracy rates using different classifiers on the Breast dataset

It is noted that classification accuracy does not increase monotonically with the increment of gene numbers. For datasets with relatively small numbers of samples, e.g. the expression datasets used in this experiment, the less number of genes provides a simpler mapping from genes to labels, which makes the classification process easier. When the number of genes increases, the classification rates may increase or decrease because of the incrementally complex co-relationship between the genes. The stability of the classification accuracy curve depends on the agreement of co-relation identification between the feature selection algorithm and the classifier. In this experiment, we concluded that the RELM is the most suitable classifier for the MIMAGA-Selection algorithm.

In summary, all four classifiers in Figures 5-7 reach the classification accuracy rates higher than 80% for all datasets, which demonstrates the robustness of the MIMAGA-Selection method. The selected small subsets of genes carry the most important information of the original datasets. The efficiency of the real-world applications, such as the cancer study, clustering and identification, can be tremendously improved.

6. Conclusion

In this work, we propose a hybrid feature selection method combining MIM and AGA and name it as MIMAGA-Selection algorithm. The MIMAGA-Selection algorithm effectively reduces the dimension of the original gene expression datasets and removes the redundancies of the data. For datasets with the number of genes up to 20,000, the MIMAGA-Selection algorithm is always capable to reduce the gene number to below 300 with reasonably high classification accuracies. The classification accuracy rates comparison with other existing feature selection algorithms shows the effectiveness of the MIMAGA-Selection algorithm. Four different classifiers, namely BP, SVM, ELM and RELM are applied to the reduced dataset. The lowest classification accuracy is around 80% which is still in the acceptable region.

The future work of this study is to improve the efficiency of the MIMAGA-Selection algorithm. While the microarray gene expression data grows exponentially in size, it takes a relatively long time for an iterative feature selection algorithm, such as MIMAGA-Selection, to reduce the dataset into small size. One possible solution is to integrate the MIMAGA-Selection into a cloud platform [44, 45]. The cloud computing provides the parallel and distributed running environment. The time complexity of the MIMAGA-Selection algorithm can be largely improved on cloud platforms.

Acknowledgements

This study is supported by National Natural Science Foundation of China (No. 61272315, No. 61303183, No.61402417, No. 60842009, and No. 60905034), Zhejiang Provincial Natural Science Foundation (No. Y1110342, No. LY15F020037, and No. Y1080950) and Zhejiang Provincial Science and Technology Department of International Cooperation Project (No. 2012C24030). It is also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET).

References

[1] Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. Annual review of biomedical engineering, 4(1), 129-153.

[2] Li, S., & Li, D. (2008). DNA Microarray Technology. In DNA Microarray Technology And Data Analysis In Cancer Research (pp. 1-9).

[3] Yu, Z., Chen, H., You, J., Wong, H. S., Liu, J., Li, L., & Han, G. (2014). Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 11(4), 727-740.

[4] Yu, Z., Chen, H., You, J., Han, G., & Li, L. (2013). Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Biomolecular Data. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 10(3), 657-670.

[5] Yu, Z., Li, L., You, J., Wong, H. S., & Han, G. (2012). SC^3: Triple Spectral Clustering-Based Consensus Clustering Framework for Class Discovery from Cancer

Gene Expression Profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(6), 1751-1765.

[6] Brazma, A., & Vilo, J. (2000). Gene expression data analysis. FEBS letters, 480(1), 17-24.

[7] Sherlock, G. (2000). Analysis of large-scale gene expression data. Current opinion in immunology, 12(2), 201-205.

[8] Torkkola, K. (2003). Feature extraction by non parametric mutual information maximization. The Journal of Machine Learning Research, 3, 1415-1438.

[9] Jakobović, D., & Golub, M. (1999). Adaptive genetic algorithm. CIT. Journal of computing and information technology, 7(3), 229-235.

[10] Yu, Z., Li, L., Liu, J., & Han, G. (2015). Hybrid Adaptive Classifier Ensemble. Cybernetics, IEEE Transactions on, 45(2), 177-190.

[11] Yu, Z., Chen, H., Liu, J., You, J., Leung, H., & Han, G. (2015). Hybrid k-Nearest Neighbor Classifier. Cybernetics, IEEE Transactions on, 46(6), 1263-1275.

[12] Kabir, M. M., Shahjahan, M., & Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. Neurocomputing, 74(17), 2914-2928.

[13] Kabir, M. M., Islam, M. M., & Murase, K. (2010). A new wrapper feature selection approach using neural network. Neurocomputing, 73(16), 3273-3283.

[14] Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection. Soft Computing, 12(2), 111-120.

[15] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[16] Huang, D., & Chow, T. W. (2005). Effective feature selection scheme using mutual information. Neurocomputing, 63, 325-343.

[17] Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. Information Sciences, 179(19), 3218-3229.

[18] François, D., Rossi, F., Wertz, V., & Verleysen, M. (2007). Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing, 70(7), 1276-1288.

[19] Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: a mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371-6385.

[20] Sha-Sha, W., Hui-Juan, L., Wei, J., & Chao, L. (2014). A Construction Method of Gene Expression Data Based on Information Gain and Extreme Learning Machine Classifier on Cloud Platform. computing, 7(2).

[21] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello Coello, C. (2014). A survey of multiobjective evolutionary algorithms for data mining: Part I. Evolutionary Computation, IEEE Transactions on, 18(1), 4-19.

[22] Eren, K., Deveci, M., Küçüktunç, O., & Çatalyürek, Ü. V. (2013). A comparative analysis of biclustering algorithms for gene expression data. Briefings in bioinformatics, 14(3), 279-292.

[23] Ahmad, F., Isa, N. A. M., Hussain, Z., & Osman, M. K. (2013). Intelligent medical disease diagnosis using improved hybrid genetic algorithm-multilayer perceptron network. Journal of medical systems, 37(2), 1-8.

[24] Yun, Y., Chung, H., & Moon, C. (2013). Hybrid genetic algorithm approach for precedence-constrained sequencing problem. Computers & Industrial Engineering, 65(1), 137-147.

[25] Silva, O., Gabriel, R., de Souza Ribeiro, M. W., & Rodrigues do Amaral, L. (2013, June). Building high level knowledge from high dimensionality biological dataset (NCI60) using Genetic Algorithms and feature selection strategies. In Evolutionary Computation (CEC), 2013 IEEE Congress on (pp. 578-583). IEEE.

[26] Diaz, J. M., Pinon, R. C., & Solano, G. (2014, July). Lung cancer classification using genetic algorithm to optimize prediction models. In Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on (pp. 1-6). IEEE.

[27] Bagyamani, J., K. Thangavel, and R. Rathipriya. (2013). Biclustering of gene expression data based on hybrid genetic algorithm. International Journal of Data Mining, Modelling and Management 5.4 (2013): 333-350.

[28] Yang, H., Yi, J., Zhao, J., & Dong, Z. (2013). Extreme learning machine based genetic algorithm and its application in power system economic dispatch. Neurocomputing, 102, 154-162.

[29] Srinivas, M., & Patnaik, L. M. (1994). Adaptive probabilities of crossover and mutation in genetic algorithms. Systems, Man and Cybernetics, IEEE Transactions on, 24(4), 656-667.

[30] Hinterding, R., Michalewicz, Z., & Peachey, T. C. (1996). Self-adaptive genetic algorithm for numeric functions. In Parallel Problem Solving from Nature—PPSN IV (pp. 420-429). Springer Berlin Heidelberg.

[31] Jakobović, D., & Golub, M. (1999). Adaptive genetic algorithm. CIT. Journal of computing and information technology, 7(3), 229-235.

[32] Qu, H., Xing, K., & Alexander, T. (2013). An improved genetic algorithm with co-evolutionary strategy for global path planning of multiple mobile robots. Neurocomputing, 120, 509-517.

[33] Chan, F. T., Chung, S. H., & Chan, P. L. Y. (2005). An adaptive genetic algorithm with dominated genes for distributed scheduling problems. Expert Systems with Applications, 29(2), 364-371.

[34] Kim, K., Yun, Y., Yoon, J., Gen, M., & Yamazaki, G. (2005). Hybrid genetic algorithm with adaptive abilities for resource-constrained multiple project scheduling. Computers in industry, 56(2), 143-160.

[35] Wang, L., & Tang, D. B. (2011). An improved adaptive genetic algorithm based on hormone modulation mechanism for job-shop scheduling problem. Expert Systems with Applications, 38(6), 7243-7250.

[36] Chen, R., Liang, C. Y., Hong, W. C., & Gu, D. X. (2015). Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. Applied Soft Computing, 26, 435-443.

[37] Gu, B., Sheng, V. S., Wang, Z., Ho, D., Osman, S., & Li, S. (2015). Incremental learning for ν-support vector regression. Neural Networks, 67, 140-150.

[38] Montana, D. J., & Davis, L. (1989, August). Training Feedforward Neural Networks Using Genetic Algorithms. In IJCAI (Vol. 89, pp. 762-767).

[39] Gu, B., Sheng, V. S., Tay, K. Y., Romano, W., & Li, S. (2015). Incremental support vector learning for ordinal regression. Neural Networks and Learning Systems, IEEE Transactions on, 26(7), 1403-1416.

[40] Gu, B., & Sheng, V. S. (2016). A robust regularization path algorithm for ν-support vector classification. IEEE Transactions on Neural Networks and Learning Systems.

[41] Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. The Journal of Machine Learning Research, 3, 1371-1382.

[42] Somol, P., Pudil, P., Novovičová, J., & Paclık, P. (1999). Adaptive floating search methods in feature selection. Pattern recognition letters, 20(11), 1157-1163.

[43] Iosifidis, A., Tefas, A., & Pitas, I. (2014). Regularized extreme learning machine for multi-view semi-supervised action recognition. Neurocomputing, 145, 250-262.

[44] Xia, Z., Wang, X., Sun, X., & Wang, Q. (2015). A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. IEEE Transactions on Parallel and Distributed Systems, 20(2), 340-352.

[45] Ren, Y., Shen, J., Wang, J., Han, J., & Lee, S. (2015). Mutual verifiable provable data auditing in public cloud storage. Journal of Internet Technology, 16(2), 317-323.

Author Biography

Huijuan Lu received her Ph.D. and B.S. from China University of Mining & Technology, the M.S. from Zhejiang University. Now she is the Professor of China Jiliang University. She is the executive director of CCF and the member of China cloud computing Expert Committee. She is principally engaged in cloud computing, pattern recognition, bioinformatics, data mining.

Junying Chen is currently a graduate student in College of Information Engineering, China Jiliang University. He obtains B.S. from Northeast Petroleum University, China. His research interests include Cloud Computing, Database Management and Machine Learning.

Dr. Ke Yan completed both the Bachelor's and Ph.D. degree in National University of Singapore (NUS). He received his Ph.D. certificate in computer science in 2012 under the supervision of Dr. Ho-Lun Cheng. During the years between 2013 and 2014, he was a post-doctoral researcher in Masdar Institute of Science and Technology in Abu Dhabi, UAE. Currently, he serves as a lecturer in China Jiliang University, Hangzhou, China. His main research field includes computer graphics, computational geometry, data mining and machine learning.

Qun Jin is currently a tenured full professor and the chair of the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Japan. He has been engaged extensively in research works in the fields of computer science, information systems, and social and human informatics. He seeks to exploit the rich interdependence between theory and practice in his work with interdisciplinary and integrated approaches. Dr. Jin has published more than 200 refereed papers in the academic journals, such as ACM Transactions on Intelligent Systems and Technology, IEEE Transactions on Learning Technologies, and Information Sciences (Elsevier), and international conference proceedings in the related research areas. His recent research interests cover human-centric ubiquitous computing, behavior and cognitive informatics, data analytics and big data security, personal analytics and individual modeling, cyber-enabled applications in e-learning and e-health, and computing for well-being. He is a member of IEEE, IEEE CS, and ACM, USA, IEICE, IPSJ, and JSAI, Japan, and CCF, China.

Yu Xue was born in 1981. He is a member of IEEE (92058890), ACM (2270255), and CCF (E200029023M).He received the Ph.D. degree fromCollege of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, China, in 2013. He is a lecturer in the School of Computer and Software, Nanjing University of Information Science and Technology. He has published nearly twenty journal and conference papers. His research interests include computational intelligence, internet of things and electronic countermeasure.

Zhigang Gao received the Ph.D. degree from the College of Computer Science, Zhejiang University, Hangzhou, China in 2008. He is a teacher in the College of Computer Science, Hangzhou Dianzi University, Hangzhou, China. His current research interests are pervasive computing, Cyber-Physical Systems, and automotive electronic systems.