MGRFE: multilayer recursive feature elimination based on embedded genetic algorithm for cancer classification

Cheng Peng , Xinyu Wu , Wen Yuan , Xinran Zhang , Yu Zhang and Ying Li

Abstract—**Background:** Microarray gene expression data has become a topic of great interest for cancer classification and further research in the field of bioinformatics. But due to the "large p, small n" paradigm of limited bio-samples and high-dimensional data, gene selection becomes a demanding task which aims at selecting minimal discriminatory genes associated closely with the phenotypes. Feature or gene selection is a still challenging problem for its NP time complexity and most of the existed feature selection algorithms utilize heuristic rules.
Results: A multilayer recursive feature elimination algorithm based on embedded variable chromosome length genetic algorithm, MGRFE, is proposed, which aims at selecting the gene combination with minimal size and maximal information. Based on 19 benchmark microarray datasets including multiclass and imbalanced datasets, MGRFE outperforms most of popular feature selection algorithms with better cancer classification accuracy and smaller selected gene number. Moreover, the selected genes by MGRFE also have closely biological relevant to cancer phenotypes.
Conclusion: Conclusion: MGRFE selects a relative small feature subset with excellent cancer classification performance through comprehensive comparison with mostly state-of-the-art methods on large scale datasets. The source code and all 19 datasets used in this paper are available at http://ml.jlu.edu.cn//MGRFE.

Index Terms—Gene selection, Genetic algorithm, Recursive feature elimination, Microarray data, Cancer classification.

1 Introduction

The chief challenge in bioinformatics is about the "large p small n" paradigm [1], on account of ever-increasing high-dimension data and limited available experimental samples [2]. It is difficult to acquire sufficient and appropriate bio-samples due to high expense of microarray sample collection and other various factors [3], and especially for gene expression data the sample number is usually small compared with several thousands to tens of thousands of genes. Therefore, selecting the most significant features from a large feature range turns out to be a demanding job [4]. Feature selection is commonly used in pre-processing and data analysis for high dimensional data problems. At the aim of reducing the dimensionality of datasets and improving the interpretability and extensibility of feasible models, feature selection attempts to eliminate irrelevant and redundant features [5], and only extract features most relative with the phenotypes and yielding best classification accuracy.

In order to find more pivotal subsets of informative features which ideally represent the target genes dataset, researchers proposed enormous feature selection algorithms. Without well performed methods of feature selection, several unexpected problems which are associated with poor scalability and incomplete information will appear. From the perspective of execution time on processors, the computational complexity of feature selection belongs to NP-hard problems [6]. On the basis of the determination-process of choosing features for classification, all feature selection methods can be roughly divided into three categories: filter, wrapper and hybrid techniques [7].

Filter algorithms generally evaluate and select attribute indexes according to the inherent characteristics of datasets, and then ranks all features so as to form an optimal subset of genetic features. Up to now, lots of filter algorithms have been designed,

- *Cheng Peng, Xinyu Wu, Wen Yuan, Xinran Zhang, Yu Zhang, and Ying Li are with the College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China.*
- *Ying Li is correspondence author. Email:liying@jlu.edu.cn.*

such as methods based on T-test [8], Chi2 [9], Mutual information [10] and maximal information coefficient (MIC) [11][12], Signal-to-Noise-Ratio [13] and so forth. As there is no classification algorithms involved in the filter algorithm, its computational speed is fast and it is suitable for large datasets. Whereas this kind of methods are probable to add redundant features into chosen subsets, which will lead to the inaccessibility of optimal results. In the gene expression datasets, genes in a cell interact with other ones to complete certain biological function but the filter methods select genes individually rather than select gene combinations which is another shortcoming [14]. And the top-ranked feature by filter algorithm are not always best features for classification [15].

Wrapper algorithms usually employ classification models and utilize heuristic rules to select feature subsets guided by the classification performance on used feature subsets, which are usually superior to filter algorithms but the consumption time is exponentially increased. A variety of wrapper algorithms have emerged including simulated annealing and randomized hill climbing [16], regularized random forest (RRF) [17], particle swarm optimization (PSO) [18, 19] and genetic algorithm (GA) [20]. With the rapid development of heuristic rules and evolutionary strategies which are commonly used in wrapper technique, various swarm intelligence algorithms have been applied for optimal feature selection. Subhajit et al. proposed a particle swarm optimization method based on adaptive K-nearest neighborhood (KNN) to identify a minimum meaningful feature subset in gene expression datasets [21]. Moosa et al. presented a modified artificial bee colony algorithm (ABC) to select minimum number of genes with high predictive accuracy for cancer classification [22]. Oreski et al. designed a hybrid genetic algorithm with neural networks to identify an optimum feature subset with high classification accuracy and scalability in credit risk assessment [24]. Jun et al described a guided hybrid GA to minimize the number of cost function evaluations [25]. However, all these feature selection methods based on swarm intelligence algorithms utilize the binary encoding method and lack of an explicit decline of the feature number. Meanwhile, it has been verified that just minimal number of informative genes are enough for effective diagnosis of different phenotypes in microarray gene datasets in [15, 22, 23, 29]. Using binary encoding method leads to three main shortcomings in finding optimal gene combinations in microarray gene datasets. (1), the fixed chromosome length for the encoding length must be equal to the gene range to represent all of the genes which leads to impossible of explicit decline of gene number and unnecessary space occupation when there are only several 1s among lots of 0s. (2), the different amounts of actual existed genes in different individuals. For there are different amounts of 1s, the actual number of genes vary in different individuals and can't be controlled precisely and many irrelevant genes may exist. (3), high time cost to generate the minimal informative gene combination. The size of optimal gene combinations in most of the datasets are below 10, but there are several thousands to tens of thousands of genes in each datasets, which result in less probability and high time cost for the evolution based feature selection algorithms without explicit decline of features to generate the optimal minimal gene combination. Recursive feature elimination (RFE) is a widely used strategy that give a explicit recursive reduce to the features by removing features with least weights [28, 29, 30, 65].

Hybrid algorithms are a combination of filter and wrapper strategies [31], which apply filter algorithms to narrow the searching space and remove irrelevant features, meanwhile, utilize wrapper algorithm on pre-selected subsets so as to accomplish optimal selection. Thus, hybrid algorithm can take both advantages of filter and wrapper techniques, and our work is designed as a hybrid algorithm.

MGRFE, a modified multilayer recursive feature elimination method with embedded genetic algorithm, is proposed, which combines the advantages of both evolution calculation of GA and the explicit feature elimination of RFE to achieve minimum discriminatory genes with optimal classification performance. Two filter techniques, T-test and MIC, are used in search space reduction stage to generate candidate gene set for precise wrapper search. In order to validate the performance of proposed method, we performed experiments on 19 benchmark gene expression datasets including multi-class and imbalanced datasets and then compared our results with 20 representative works using same datasets. The experiment and comparison results suggest that our method far outweighs most of popular feature selection algorithms, with apparently reduced feature number and better classification accuracy. Furthermore, the particular biomedical relevances of selected genes to the phenotypes in datasets have been verified. The whole work in this study is in Fig. 1.

2 Materials and Methods

Materials

The 19 benchmark gene expression datasets in the problem of gene selection for cancer classification [15][21] are used to validate our proposed method. The

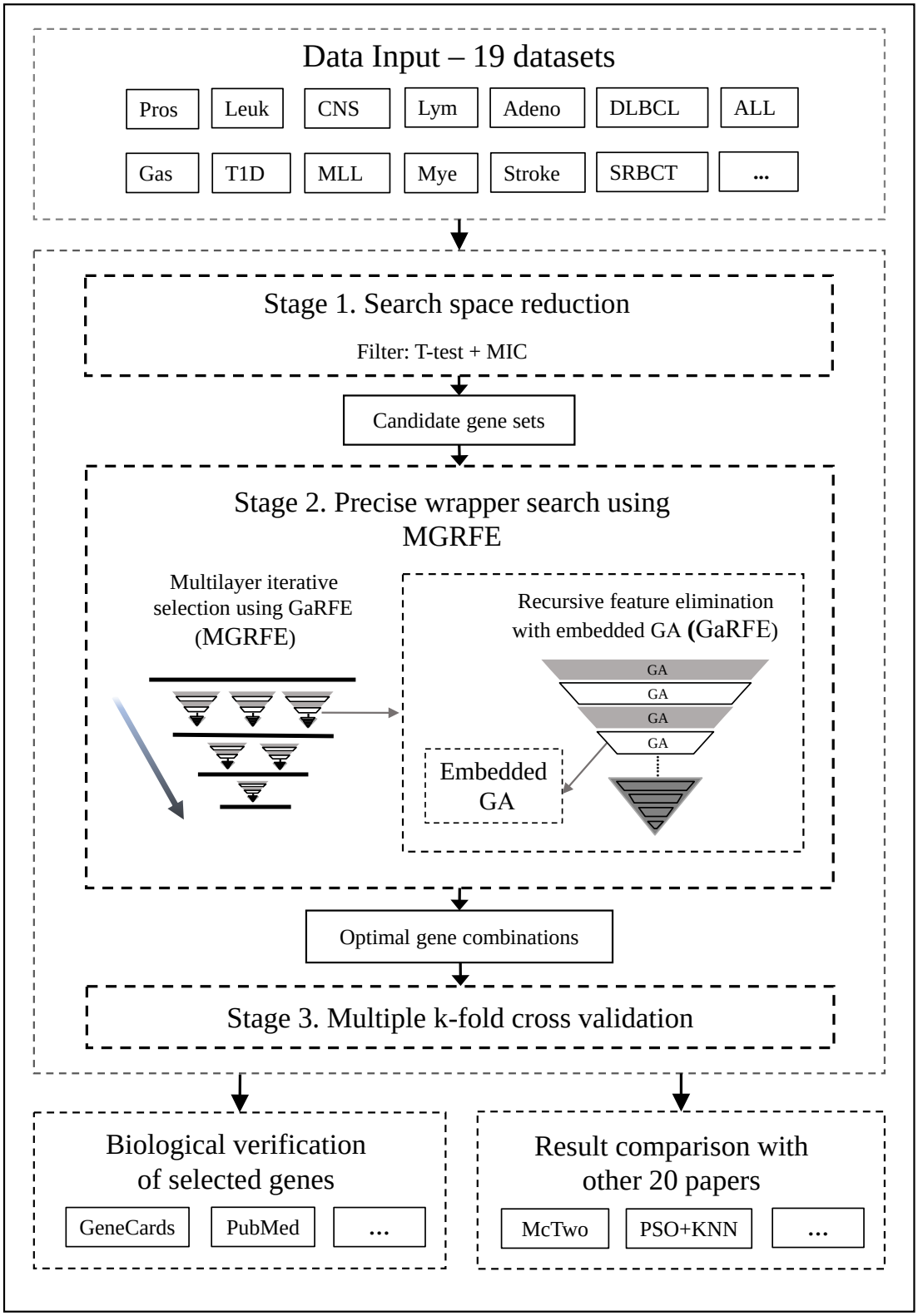

Fig. 1. The flowchart of whole work in this study

all 19 datasets include binary, multi-class, balanced and imbalanced datasets, are divided into two large Datasets. It should be mentioned that the widely used benchmark dataset Leukemia is tested in both [15] and [21] and is named as Leuk and ALL_AML respectively but they are the same dataset in actual.

Dataset One consists of 17 binary classification datasets used in [15], which includes DLBCL [32], Prostate [33], ALL(divided into four parts based on different phenotype) [34], CNS [35], Lymphoma [36], Adenoma [37], Colon [38], Leukaemia [39], Myeloma [40], Gastric [41], Gastric1/Gastric2 [42], T1D [43] and Stroke [44]. It should be noted that DLBCL, Colon, Leukaemia, Myeloma, ALL1-4, and CNS datasets are imbalanced. The brief description of the 17 datasets is in Table 1.

Dataset Two is composed of the 3 typical benchmark datasets used in [21] including 2 multi-class datasets SRBCT [51] and MLL [4] and one binary dataset ALL_AML. The brief description of the 3 datasets is shown in Table 2. There have been lots of experiments focusing at gene selection for cancer classification problem performed on these 3 datasets in [46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61]. We will offer detailed performance comparison between all these methods and the proposed method in later results section.

Classification performance measurements

According to the previous researches [62, 67, 68, 69], the accuracy (Acc) is the most widely used evaluation metric for both binary and multi-class classification tasks and thus it is regarded the main performance metric in comparison between different methods. For the 17 binary datasets including many imbalanced datasets, this study also employed another four wildly used evaluation metrics which are Sn, Sp, Avc, and MCC rather than only use Acc to measure the performance of the proposed method and handle data imbalance problem [62, 63, 64, 65, 66]. In binary classification task, the prediction outcome is often presented in a confusion matrix in Table 3 and the definitions of TP, TN, FP, FN are also shown in it. Sensitivity abbreviated to Sn indicates the proportion of correct prediction on positive samples. Specificity (Sp) is used to measure the fraction of negative samples that are correctly classified. The model's overall accuracy is Acc and average accuracy of two classes is defined as Avc. The last metric Matthews correlation coefficient (MCC) can capture all elements in the confusion matrix other than Acc and is a balanced and comprehension metric. The calculation formula of 5 metrics is in (1). In the 2 multi-class datasets, for unity and convenience, only accuracy (Acc) is used because the comparison papers all just calculated this metric.

$$\begin{aligned} Sn &= \frac{TP}{TP+FN}, \quad Sp = \frac{TN}{TN+FP}, \\ Acc &= \frac{TP+TN}{P+N}, \quad Avc = \frac{Sn+Sp}{2}, \\ MCC &= \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{aligned} \tag{1}$$

Method

The proposed method is divided into three stages as shown in Fig. 2 to find minimal discriminatory gene combinations from several thousands of genes in the microarray gene datasets.

(1) Stage 1: search space reduction using filter techniques.
(2) Stage 2: precise wrapper search using MGRFE including two key processes of GaRFE and GA.
(3) Stage 3: multiple k-fold cross validation to select final gene combination.

Stage 1: Search space reduction

The aim of the Stage 1 is to decrease the amount of genes and offer candidate gene set for later precise wrapper search stage. Two filter methods, T-test and MIC, are used in this stage to generate candidate gene set. T-test takes both the mean value and variance of the data classes into consideration in its formula and is widely used to determine if two sets of data are significantly different from each other. MIC applies mutual information to calculate the association strength between two variables having linear or nonlinear relations. Firstly, perform T-test on all genes and give them a ascending sort according to their related p-values, then only the top-ranked 1000 genes will be preserved. Secondly, carry out MIC calculation on the preserved genes and re-sort them according to their MIC values, then the candidate gene set is generated from the top 500 genes with MIC values. With regards to the 2 multi-class datasets, candidate gene set is generated just basing on the descending order of MIC values of all genes for multivariate T-test can't be performed directly.

Stage 2: Precise wrapper search using MGRFE

Stage 2 is the characteristic part of our proposed algorithm which searches the input of candidate gene set from the Stage 1 and outputs optimal gene combinations for further selection in the Stage 3. MGRFE is a multi-layer iterative feature selection method and its selection unit at each layer is a GaRFE process. GaRFE,

TABLE 1
Summary of the 17 binary classification datasets used in Dataset One from [15]

ID	Dataset	Samples	Features	Summary
1	DLBCL[1]	77	7129	DLBCL patients (58) and follicular lymphoma (19)
2	Pros(Prostate)[1]	102	12625	prostate (52) and non-prostate (50)
3	Colon[2]	62	2000	tumour (40) and normal (22)
4	Leuk(Leukaemia)[2]	72	7129	ALL (47) and AML (25)
5	Mye(Myeloma)[3]	173	12625	presence (137) and absence (36) of focallesions of bone
6	ALL1[1]	128	12625	B-cell (95) and T-cell (33)
7	ALL2[1]	100	12625	patients that did (65) and did not (35) relapse
8	ALL3[1]	125	12625	with (24) and without (101) multidrug resistance
9	ALL4[1]	93	12625	with (26) and without (67) the t(9;22) chromosome translocation
10	CNS[1]	60	7129	medulloblastoma survivors (39) and treatment failures (21)
11	Lym(Lymphoma)[1]	45	4026	germinalcentre (22) and activated B-like DLBCL (23)
12	Adeno(Adenoma)[1]	36	7457	colon adenocarcinoma (18) and normal (18)
13	Gas(Gastric)[3]	65	22645	tumors (29) and non-malignants (36)
14	Gas1(Gastric1)[3]	144	22283	non-cardia (72) of gastric and normal (72)
15	Gas2(Gastric2)[3]	124	22283	cardia (62) of gastric and normal (62)
16	T1D[3]	101	54675	T1D (57) and healthy control (44)
17	Stroke[3]	40	54675	ischemic stroke (20) and control (20)

In this table of Dataset One and the next table of Dataset Two, "Sample" and "Features" indicate the total sample number and feature number and "Summary" column describes the sample classes and the related sample number in parenthesis. This table is just as the description table of 17 datasets in [15].

[1] These datasets are retrieved from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[2] Colon and Leuk datasets are downloaded from the R/Bioconductor packages colonCA and golubEsets respectively.

[3] These datasets are downloaded form https://www.ncbi.nlm.nih.gov/geo/

TABLE 2
Summary of the 3 classification datasets used in Dataset Two from [21]

ID	Dataset	Classes	Samples	Features	Summary
1	SRBCT[1]	4	88	2308	EWS (29), NHL (11), NB (18) and RMS (25)
2	ALL_AML[2]	2	72	7129	ALL (47) and AML (25)
3	MLL[3]	3	72	12582	ALL (24), MLL (20) and AML (28)

[1] SRBCT dataset is downloaded from http://research.nhgri.nih.gov/microarray/Supplement/, which includes 88 samples totally, but 5 of them are irrelevant and thus only 83 samples are used.

[2] ALL_AML in Dataset Two and Leuk in Dataset One are same datasets in actual.

[3] MLL dataset is retrieved from http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?paper_id=63

TABLE 3
Confusion matrix for binary classification and explain for TP, FP, TN, FN

		Actual class	
		Positive ($P = TP + FN$)	Negative ($N = TN + FP$)
Predicted class	Positive($P^{'} = TP + FP$)	True Positive (TP)	False Positive (FP)
	Negative($N^{'} = FN + TN$)	False Negative (FN)	True Negative (TN)

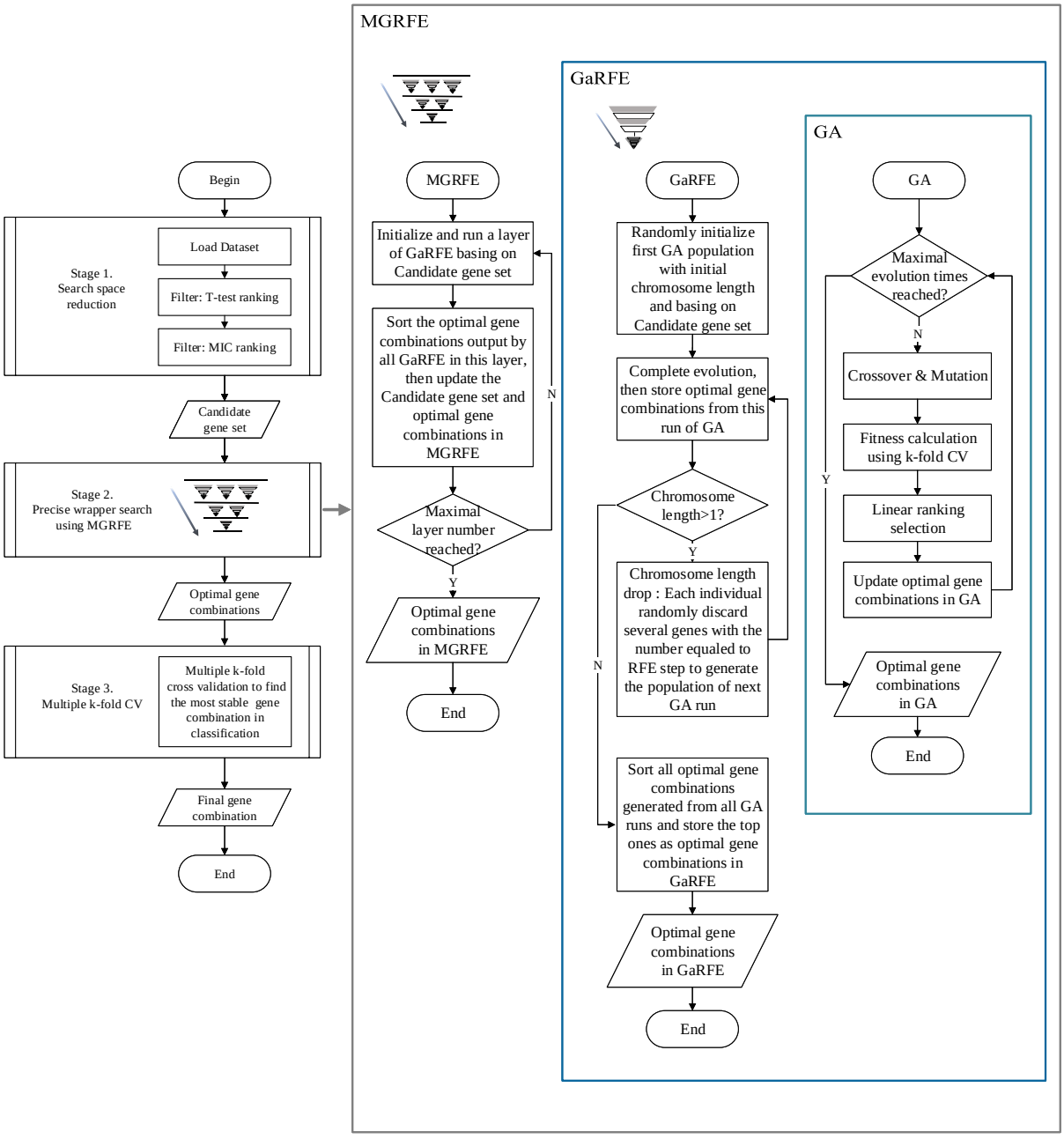

Fig. 2. The flowchart of MGRFE which is divided into 3 stages: search space reduction, precise wrapper search using MGRFE and multiple k-fold CV. The stage 2 is the core of MGRFE, which includes two key processes: GaRFE and embedded modified GA.

the inverted triangle in MGRFE as shown in Fig. 1, is the recursive feature elimination process with every stair being embedded genetic algorithm. Embedded GA is the modified genetic algorithm using integer encoding method and variable length chromosome. The key of MGRFE is GaRFE at each layer, where embedded modified GA is responsible for generating optimal gene combinations and RFE is responsible for cutting down the gene number. Therefore our method can find gene combinations with both significantly re-

duced sizes and excellent classification performances.

Embedded genetic algorithm: In our method, the embedded modified GA using variable length integer encoding chromosome is embedded in the recursive feature elimination process as each stair in the inverted triangle of GaRFE. The embedded GA includes following steps. To begin with, generate some individuals representing gene combinations with same size. Then, perform fitness calculation and genetic operators including mutation, crossover and selection until stopping criteria is satisfied. In the end, return the best individuals which represents best gene combinations to GaRFE.

To embed GA in the RFE process and achieve the goal of finding minimal informative genes, we made some modification on the original GA. The embedded GA uses variable length integer encoding method for the chromosome in individuals and each individual has a set of integers representing different genes which make up a gene combination. In every run of GA, the gene combinations represented by different individuals all have a fixed size. Between two runs of GA in the RFE process, every individual decreases a same number of genes from its chromosome.

The mutation and crossover operators for generating new individuals should also be adjusted to adapt to the variable length integer encoding method. One main problem to be concerned is avoiding duplicated genes in every individual which lead to the decline of actual existed genes. Based on our encoding method, randomly change some genes to others is the mutation operation. It should be ensured that new genes don't exist in this individual previously to avoid repetitions. Crossover in embedded GA also uses single point crossover which is the most widely used crossover method in binary encoding. Specifically, a random position is selected in the chromosome, and two parent individuals split themselves at this crossover point and then exchange chromosome tails to generate children individuals. After crossover, replace the potential duplicates genes in the children individuals with other genes from their parents to avoid decreasing of gene number.

Fitness (F) of an individual is defined as follows:

$$F = \begin{cases} Acc & \text{, balanced dataset} \\ \alpha Acc + (1-\alpha)Avc & \text{, imbalanced dataset.} \end{cases} \quad (2)$$

In (2) Acc represents the average accuracy from 5-fold cross validation and α is an adjustment coefficient to dual with the imbalance problem of datasets. For imbalanced datasets, fitness defined as $\alpha Acc + (1-\alpha)Avc$ can adjust the tend of predicting samples as abundant classes [71] for $Avc = (Sn + Sp)/2$ takes the correct prediction proportion of both two sample classes into consideration, and this fitness definition can be regarded as a new way to help tackle the data imbalanced problem. For balanced datasets, F can be directly calculated from Acc. We uses 5-fold cross validation to calculate the fitness value and the employed classifier is Naive Bayes classifier (NBayes) [70]. The sorting between different individuals or gene combinations is based on two metrics, fitness and gene number. For two individuals with different fitness values, the one with higher fitness is more superior and for two individuals having same fitness values, the one with smaller gene number is more superior. Individuals are exactly gene combinations in the proposed method, and MGRFE and GaRFE also use this sorting rule to rank different gene combinations.

Selection process in embedded GA uses linear ranking selection method [72] which simply ranks all individuals according to their related fitness values and select the top individuals to form the next generation. The stopping criteria is iteration times which are set from 1 to 3. In each iteration, perform mutation, crossover, fitness calculation and selection and then update the fittest individuals which are the optimal gene combinations. When the maximal iteration time is reached, embedded GA returns its overall optimal gene combinations to the GaRFE.

Recursive feature elimination with embedded modified GA: GaRFE is designed as an explicit recursive feature elimination of genes with embedded modified GA to find minimal discriminatory gene combinations. First, randomly generate the initial GA population basing on certain candidate gene set and chromosome length. Then, perform chromosome length drop and GA iteration in turns until the chromosome length in GA drop to 1. Finally, sort all optimal gene combinations returned by all GA runs and then return the overall top-ranked gene combinations to the MGRFE. Chromosome length drop means that every individual in current GA population randomly discards the same number of genes to generate the new GA population for the next run. The number of discarded genes is set from 1 to 3 according to current chromosome length. Larger decline step set for larger chromosome length to avoid time cost and smaller decline step set for smaller chromosome length to do precisely searching.

Multilayer iterative selection: MGRFE employs multilayer iterative feature selection method with the selection unit at each layer as GaRFE. GaRFE at every iteration layer uses the current candidate gene set and returns its overall optimal gene combinations, and then the candidate gene set is reduced and used for the next layer of iteration selection. The candidate gene set used by the first layer of MGRFE is from

the search space reduction stage. After each iteration, all gene combinations in MGRFE will be sorted and the top-ranked combinations will be taken as the updated reduced candidate gene set. After specified layers of iteration, MGRFE sorts all returned optimal gene combinations and outputs the top-ranked gene combinations for Stage 3 to make a further validation.

Stage 3: Multiple k-fold CV to select most stable gene combination

This Stage 3 aims at finding the most excellent gene combination with best classification performance and lowest variance among different CV results. K-fold CV is performed in calculating the fitness of the GA individual. Multiple k-fold CV based on different random seeds is performed here to further validate and select the overall optimal gene combination.

3 RESULTS

The section presents the performance comparison with other methods on 19 benchmark datasets divided into two large Datasets. In addition, the biological verification of the selected genes is also discussed.

Results on Dataset One

The results of MGRFE for 17 binary datasets is in Table 4, including five evaluation metrics calculated by 5-fold CV and T-test based gene ranks. Experiment results confirmed that our method can find the minimal gene combinations with excellent phenotypes diagnosis abilities on all 17 datasets. Acc values are all above 0.9 with gene number below 10. Furthermore, in 8/17 datasets of DLBCL, Leuk, ALL1, Lym, Adeno, Gas, Gas2 and Stroke, Acc can reach 1.0 with gene number less than 5. MGRFE also shows robust stability in dealing with imbalanced dataset like DLBCL, Colon, Leuk, ALL1, ALL4 and CNS, in which Sn, Sp, Avc and MCC values are also very high and haven't been influenced by the imbalanced problem. From the T-test based gene rank positions which begin with 0, we can note that the best gene or feature subset is not always the top features in T-test and only filter algorithm can't generate the best feature combinations. But meanwhile in 5/17 datasets, the top one gene from T-test which has position number 0 appeared in the selected final gene combinations and lots of top-ranked genes are also in, therefore the filter techniques are qualified for search space reduction stage. Moreover, MGRFE achieves stable classification performance using 10 times 10-fold cross validation as shown in Fig. 3.

Comparison with other method on Dataset One

For McTwo in [15] completely tested all datasets in Dataset One, here we offer the performance comparison between McTwo and MGRFE. Table 5 shows the overall Acc and number of selected genes on all 17 datasets by MGRFE and McTwo. For more intuitively comparison, Fig. 4 offers the line chart of maximal Acc achieved by two algorithms on all datasets and MGRFE obviously outperforms McTwo with a higher Acc line.

For 5 datasets of ALL2, ALL3, ALL4, Stroke and CNS, MGRFE achieved distinctly better classification performance than McTwo with relative more genes. For more fair and specific comparison, we further listed the overall accuracies when gene numbers of MGRFE are equal to McTwo as shown in Table 6. The results show that MGRFE still outperforms McTwo when MGRFE uses the same gene number as McTwo used. But the Acc values by using these gene number fall behind our optimal Acc values in these datasets by a large margin and therefore MGRFE selected little more genes to achieve the optimal results.

Results on Dataset Two

Here presents the results of MGRFE for 3 typical benchmark datasets including two multiclass datasets. In datasets of SRBCT, ALL_AML and MLL, 5, 2 and 3 genes were selected respectively and the overall maximal Acc is all 1.0. Fig. 5 offers three instances of GaRFE process at the first layer of MGRFE on these datasets, and from which we can discover that Acc values quickly rise to 1.0 in the initial GA iterations in GaRFE and only begin to drop when the gene number is significantly reduced. 10 times 10-fold cross validations are carried to further validate the final selected gene combinations on Dataset Two as shown in Fig. 6. The multiple Acc on SRBCT, ALL_AML and MLL are 98.8%, 98.3%, and 99.7% respectively, which shows that MGRFE has high classification stability.

Comparison with other methods on Dataset Two

The performance comparison according to metrics of Acc and gene number with the mostly used feature selection methods on the 3 benchmark datasets as shown in Table 7, Table 8 and Table 9.

In SRBCT dataset, Khan et al. [51] applied artificial neural network (ANN) and used 96 genes to achieve 100% Acc; Tibshirani et al. [58] used nearest shrunken centroid based method and achieved 100% Acc by 43 genes; Fu and Fu-Liu. [48] used SVM-RFE and achieved 100% Acc by 19 genes; Pal et al. [55] applied multi layered perceptron and non-Euclidean relational fuzzy c-means clustering and find 7 genes important

TABLE 4
Results of MGRFE on 17 datasets in Dataset One

Dataset	Pos/Neg	Genes	Sn	Sp	Acc	Avc	MCC	T-test based gene ranks
DLBCL	58/19	**3**	1.0	1.0	**1.0**	1.0	1.0	[12, 38, 53]
Pros	52/50	**4**	0.980	0.982	**0.981**	0.981	0.963	[0, 14, 73, 693]
Colon	40/22	**6**	1.0	0.960	**0.985**	0.980	0.969	[14, 57, 175, 224, 239, 494]
Leuk	47/25	**2**	1.0	1.0	**1.0**	1.0	1.0	[3, 6]
Mye	137/36	**7**	0.963	0.839	**0.937**	0.901	0.816	[2, 14, 82, 142, 377, 403, 568]
ALL1	95/33	**1**	1.0	1.0	**1.0**	1.0	1.0	[0]
ALL2	65/35	**8**	0.914	0.908	**0.910**	0.911	0.829	[0, 51, 77, 79, 521, 686, 736, 759]
ALL3	24/101	**8**	0.830	0.950	**0.927**	0.890	0.785	[3, 51, 74, 141, 487, 509, 714, 769]
ALL4	26/67	**6**	1.0	0.986	**0.990**	0.993	0.978	[0, 5, 38, 281, 534, 753]
CNS	39/21	**7**	1.0	1.0	**1.0**	1.0	1.0	[8, 52, 129, 130, 271, 272, 519]
Lym	22/23	**3**	1.0	1.0	**1.0**	1.0	1.0	[3, 4, 668]
Adeno	18/18	**1**	1.0	1.0	**1.0**	1.0	1.0	[467]
Gas	29/36	**3**	1.0	1.0	**1.0**	1.0	1.0	[21, 76, 305]
Gas1	72/72	**3**	0.986	0.973	**0.980**	0.980	0.961	[131, 247, 716]
Gas2	62/62	**2**	1.0	1.0	**1.0**	1.0	1.0	[37, 88]
T1D	57/44	**7**	0.911	0.912	**0.911**	0.912	0.826	[13, 24, 112, 558, 577, 679, 977]
Stroke	20/20	**4**	1.0	1.0	**1.0**	1.0	1.0	[0, 22, 128, 275]

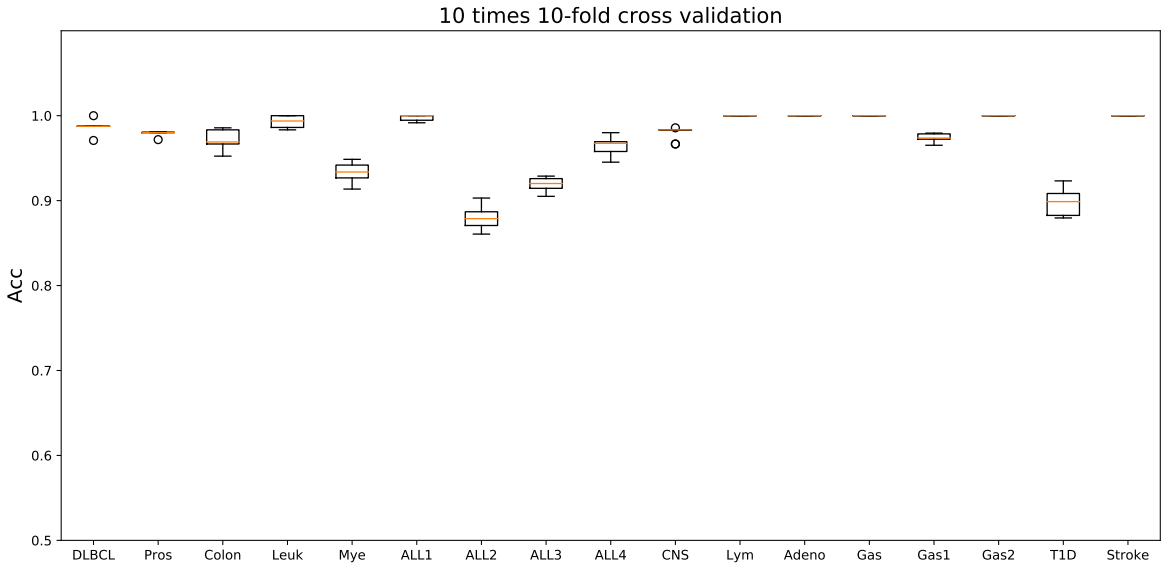

Fig. 3. The distribution of Acc values for 10 times 10-fold cross validation on the selected gene combinations of 17 datasets

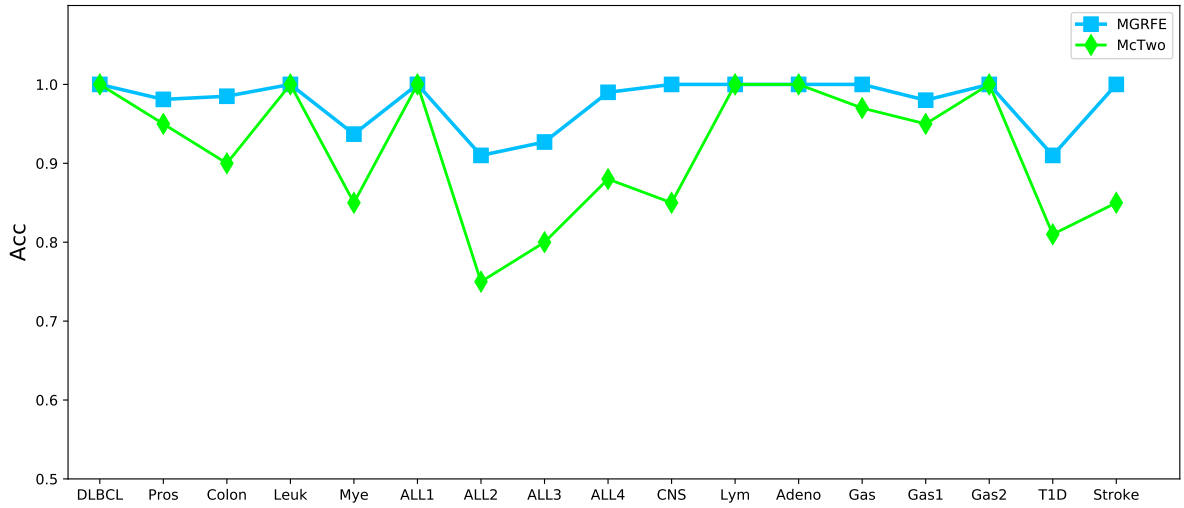

Fig. 4. The line plots of overall maximal accuracy for MGRFE and McTwo on the 17 datasets

TABLE 5
Performance comparison between McTwo and MGRFE on 17 datasets

	DLBCL	Pros	Colon	Leuk	Mye	ALL1	ALL2	ALL3	ALL4	CNS	Lym	Adeno	Gas	Gas1	Gas2	T1D	Stroke
MGRFE Acc	1	0.981	0.985	1	0.937	1	0.91	0.927	0.99	1	1	1	1	0.98	1	0.91	1
McTwo Acc	1	0.95	0.9	1	0.85	1	0.75	0.8	0.88	0.85	1	1	0.97	0.95	1	0.81	0.85
MGRFE Genes	3	3	6	2	7	1	8	8	6	7	3	1	3	3	2	7	4
McTwo Genes	4	3	6	2	7	1	2	5	2	4	4	2	3	4	2	6	1

TABLE 6
Performance comparison on 5 datasets between MGRFE and McTwo with same gene number

Dataset	Method	Genes	*Acc*
ALL2	MGRFE	2	0.760
	McTwo	2	0.75
ALL3	MGRFE	5	0.874
	McTwo	5	0.8
ALL4	MGRFE	2	0.896
	McTwo	2	0.88
CNS	MGRFE	4	0.921
	McTwo	4	0.85
Stroke	MGRFE	1	0.825
	McTwo	1	0.75

for 100% *Acc*; Mohamad et al. [54] used improved binary PSO and 6 genes were selected; Subhajit et al. [21] applied PSO and KNN and also 6 genes were selected; Moosa et al. [22] achieved 100% *Acc* with modified artificial bee colony algorithm by 5 genes; Sharma et al. [56] applied successive feature selection (SFS) with linear discriminant analysis (LDA) and nearest centroid classifier (NCC) and achieved 100% train and test *Acc*. In our experiments, combinations of 4 genes can also reach 100% train and test *Acc* in 5-fold CV, but these gene combinations didn't show classification stabilities in 10 times 10-fold CV and MGRFE didn't choose these 4 gene combinations. For SRBCT dataset, MGRFE selected 5 genes and achieved 100% 5-fold *Acc* and 98.5% 10 times 10-fold CV *Acc*.

In ALL_AML dataset, Fu and Fu-Liu. [48] achieved 100% train *Acc* by 19 genes based on SVM-RFE; Yang et al. [60] applied gene scoring technique and SVM and 4 genes were selected to achieve 98.6% *Acc* in leave one out cross validation (LOOCV); Mohamad et al. [54] selected 2 genes to reach 100% CV *Acc* based on improved binary PSO; Dashtban et al. [23] applied

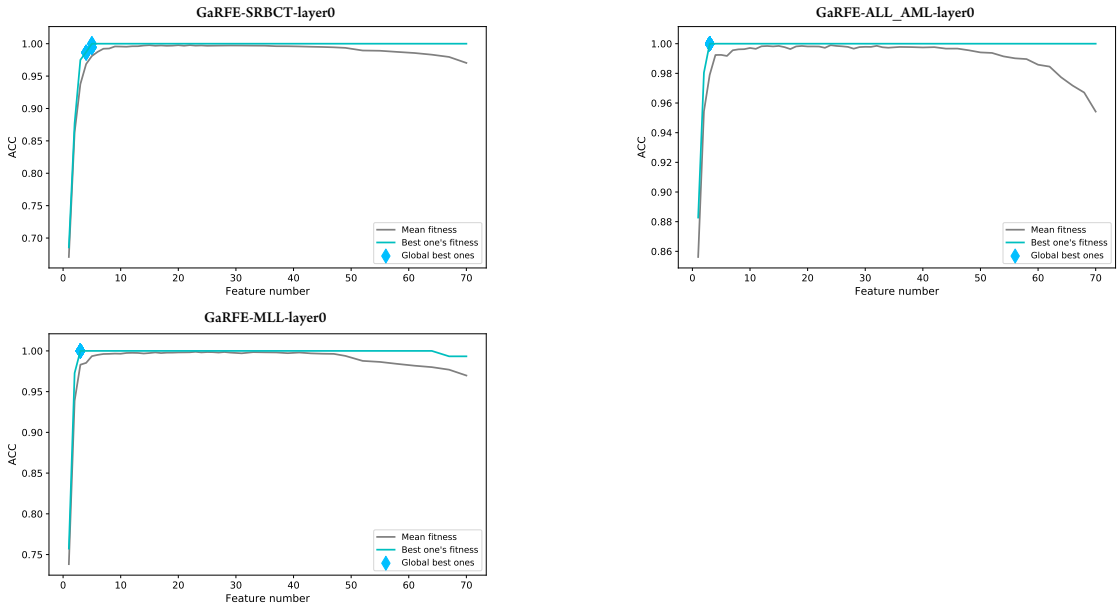

Fig. 5. Three GaRFE processes on the 3 benchmark datasets in Dataset Two

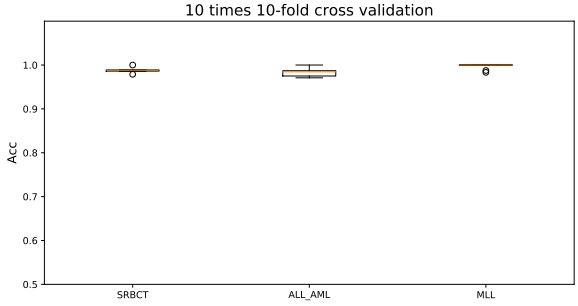

Fig. 6. The distribution of Acc values for 10 times 10-fold cross validation on the selected gene combinations of 3 benchmark datasets in Dataset Two

integer encoding GA and SVM and selected 15 genes with 100% Acc; Ruiquan et al. [15] designed a two step MIC based method and 2 genes were selected to reach 100% Acc. For ALL_AML dataset, MGRFE selected 2 genes and achieved 100% 5-fold Acc and 98.3% 10 times 10-fold CV Acc.

In MLL dataset, Sharma et al. [56] selected 4 genes with 100% train and test Acc based on SFS, LDA and NCC; Mohamad et al. [54] selected 4 genes with 100% CV Acc based on improved binary PSO; Dashtban et al. [23] applied integer encoding GA and SVM and selected 15 genes with 100% Acc; Subhajit et al. [21] applied PSO and KNN to select 4 genes with 100% train and test Acc and 92.5% CV Acc. For MLL dataset, MGRFE selected 3 genes and achieved 100% 5-fold Acc and 99.7% 10 times 10-fold CV Acc.

Biological inferences of features selected by MGRFE

This selected features by MGRFE also have closely relevances to the phenotypes in gene expression datasets. In Table 10, Table 11 and Table 12, we investigated the genes selected by MGRFE on the datasets of Leuk, Gas and ALL1 in which just 2/3/1 genes are used respectively to achieve the 5-fold CV Acc of 100%.

In Leukaemia dataset, the selected genes are CD33 and TCF3. On Pubmed there are total 2867 papers about CD33 among which 58.94% papers discussed the relevance between CD33 and leukaemia. And there are 314 papers in Pubmed confirmed the association between TCF3 and leukaemia. From GeneCards, we found that the E protein encoded by TCF3 plays a critical role in lymphopoiesis and is necessary for B and T lymphocyte. This gene is related with malignancies including acute lymphoblastic leukemia (t(1;19), with PBX1), childhood leukemia (t(19;19), with TFPT) and acute leukemia (t(12;19). In Gastric datasets, gene COL8A1, SEMA6D and LIFR are selected by MGRFE and there are 236 papers on Pubmed confirmed their relevances with cancer but just 3 papers revealed their relations with the gastric cancer. According to the excellent classification performance of these three genes for gastric cancer, they could be novel biomarker

TABLE 7
Comparison of the methods on the SRBCT dataset.

Experiments	Methods	Genes				CV Acc(%)				Train Acc(%)	Test Acc(%)
Khan et al. (2001) [51]	ANN	96				-				100	100
Tibshirani et al. (2002) [58]	NSC	43				-				100	100
Fu and Fu-Liu. (2005) [48]	SVM-RFE	19				-				100	100
		SCV		LOOCV		SCV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang et al. (2006) [60]	GS1	88	93	57	34	98	97.9	98.8	98.8	-	-
	GS2	90	99	77	96	98.1	99	98.8	100	-	-
	Chos	98	98	82	80	90.2	94.3	92.8	98.8	-	-
	F-test	90	95	89	78	98	99.2	98.8	100	-	-
Pal et al. (2007) [55]	FSMLP+NERFCM	7				-				100	100
Ji et al. (2011) [50]	PLSVIP	24				-				100	100
	PLSVEG	15				-				100	100
Mohamad et al. (2011) [54]	IBPSO	6				100				-	-
Sharma et al. (2012) [56]	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	90
	SFS+NNC	4				-				100	95
Zainuddin and Ong (2011) [61]	MSFCM+WNN	10				10CV					
						100				-	-
Li and Shu (2009) [53]	KLLE+LLE+PCA	20				-				100	100
Lee et al. (2011) [52]	AGA+KNN	14				-				100	100
Chen et al. (2014) [6]	PSODT	-				5CV					
						92.94				-	-
Subhajit et al. (2015) [21]	PSO+KNN	6				98.0159				100	100
Moosa et al. (2016) [22]	ABC	5								100	100
Dashtban et al. (2017) [23]	GA+SVM	18								100	100
This paper	**MGRFE**	**5**				**98.8**				**100**	**100**

TABLE 8
Comparison of the methods on the ALL_AML (Leukemia) dataset.

Experiments	Methods	Genes				CV Acc(%)				Train Acc(%)	Test Acc(%)
Fu and Fu-Liu (2005) [48]	SVM-RFE	19				-				100	97.06
		SCV		LOOCV		SCV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang et al. (2006) [60]	GS1	100	93	60	4	97.9	97.9	98.6	98.6	-	-
	GS2	85	98	10	25	97.1	97.4	98.6	98.6	-	-
	Chos	100	98	9	80	96.8	97	97.2	98.6	-	-
	F-test	96	99	25	33	97.4	97.5	98.6	98.6	-	-
Shen et al. (2008) [57]	Stepwise	3				-				90.83	88.14
	Pure TS	5				-				95.83	94.24
	Pure PSO	7				-				94.75	94.19
	HPSOTS	7				-				98.08	95.81
Ji et al. (2011) [50]	PLSVIP	9				-				100	100
	PLSVEG	8				-				100	100
Mohamad et al. (2011) [54]	IBPSO	2				100				-	-
Zainuddin and Ong (2011) [61]	MSFCM+WNN	10				10CV					
						98.61				-	-
Wong and Liu (2010)	Probabilistic mechanism	-				SVM	KN				
						97.38	98.21			-	-
Chandra and Gupta (2011) [47]	RNBC	-				10CV					
						RNBC	NBC	KNN			
						94.29	84.29	85.71		-	-
Ganesh Kumar et al. (2012) [49]	GSA	10				100				-	-
Subhajit et al. (2015) [21]	PSO+KNN	3				95.8868				100	97.0588
Ruiquan et al. (2016) [15]	McTwo	2								100	100
Dashtban et al. (2017) [23]	GA+SVM	15								100	100
This paper	**MGRFE**	**2**				**98.3**				**100**	**100**

TABLE 9
Comparison of the methods on the MLL dataset.

Experiments	Methods	Genes				CV Acc(%)				Train Acc(%)	Test Acc(%)
		SCV		LOOCV		SCV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang et al. (2006) [60]	GS1	29	99	97	56	94.8	95.2	97.2	97.2	-	-
	GS2	91	87	90	91	94.9	94.7	97.2	97.2	-	-
	Chos	93	89	23	44	96	95.5	97.2	95.8	-	-
	F-test	99	100	65	31	95.4	94.8	95.8	95.8	-	-
Sharma et al. (2012) [56]	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	100
	SFS+NNC	4				-				100	93
Mohamad et al. (2011) [54]	IBPSO	4				100				-	-
Chandra and Gupta (2011) [47]	RNBC	-				10CV					
						RNBC	NBC	KNN			
						87.14	80	68.57		-	-
Chen et al.(2014) [6]	PSODT					5CV					
						100				-	-
Subhajit et al. (2015) [21]	PSO+KNN	4				92.5439				100	100
This paper	**MGRFE**	**3**				**99.7**				**100**	**100**

candidates for gastric cancer biological researchers. In ALL1(acute lymphocytic leukemia) dataset, only one gene CD3D is selected by MGRFE and there are 13 papers on Pubmed revealed the relevance between CD3D and ALL. In [45], it has also been pointed out that the gene CD3D is one ideally discriminatory feature and gave the diagnostic rule that when the expression of CD3D is below certain limit. For CD3D, GeneCards explains that this gene is involved in T-cell development and signal transduction and defects in this gene will leads to severe combined immunodeficiency which is related to ALL.

Implementation notes

This study used the scipy package version 0.19.0 on the software Python version 3.6 to perform T-test. MIC calculation was implemented using the minepy package version 1.2.0. on Python version 3.6. It is noted that about 3 GaRFE in the first layer, 2 GaRFE in the second layer and 1 GaRFE in the last layer is enough for almost all datasets because the optimal gene combinations are usually found already by the GaRFE at the first layer.

4 Discussion

This paper proposed MGRFE, a novel multilayer recursive feature elimination algorithm based on embedded variable length encoding genetic algorithm, which aims at selecting minimal discriminatory genes associated closely with the phenotypes. We combined the evolutionary calculating of embedded genetic algorithm and explicit feature decline of recursive feature elimination as GaRFE, which is taken as the feature selection unit at each layer of MGRFE. The mostly used total 19 benchmark microarray datasets including multiclass and imbalanced datasets are divided into two large datasets and used to validate the proposed method and make a comprehensive comparison with other popular feature selection methods for cancer classification. Many promising results were obtained by MGRFE on these datasets. MGRFE can reaches Acc 100% within just 5 genes on 10 (52.6%) of 19 datasets, and Acc higher than 90% within 10 genes on all 19 datasets. MGRFE also shows the robustness for multi-class datasets and imbalanced datasets according to Sn, Sp, Avc and MCC metrics. Based on classification performance comparison with other 20 methods on the two large Datasets, our proposed method MGRFE is proved to be more superior than most of current popular feature selection methods for achieving better classification accuracy with smaller gene size. Furthermore, the biological function analysis using literature mining for predicted biomarkers confirmed that the selected genes by MGRFE are biologically relevant to cancer phenotypes. Therefore the minimal genes with maximal information selected by MGRFE could be novel biomarker candidates which are significant for related phenotypes study. Moreover, for clinical microbiology applications using microarray technology, MGRFE can contribute to develop potential simplified diagnosis of the cancer subgroups by designing simplified microarray genechip basing on the minimal discriminatory genes selected by MGRFE, which will cut down the cost of medical diagnoses. MGRFE can represent a complementary feature selection algorithm for high-dimensional biodata analysis and is significant for cancer diagnosis and further

TABLE 10
Literature Mining for Predicted Biomarkers for Leukaemia in PubMed

Probeset ID	Gene symbol	PubMed hits for gene of interest	PubMed hits for gene of interest and leukaemia[1](Ratio1[*])
M23197_at	CD33 Molecule(CD33)	2867	1690(58.94%)
M31523_at	Transcription Factor 3(TCF3)	5280	314(5.94%)

[1] gene of interest [All Fields] AND (leukaemia[All Fields]).
[*] Ratio1 = #(gene of leukaemia related literatures)/#(gene of interest literatures).

TABLE 11
Literature Mining for Predicted Biomarkers for Gastric Cancer in PubMed

Probeset ID	Gene symbol	PubMed hits for gene of interest	PubMed hits for gene of interest and cancer[1] (Ratio1[*])	PubMed hits for gene of interest and gastric cancer[2](Ratio2[**])
226237_at	collagen type VIII alpha 1 chain(COL8A1)	58	11(18.96%)	1(9.09%)
226492_at	semaphorin 6D(SEMA6D)	38	13(34.21%)	1(7.69%)
227771_at	leukemia inhibitory factor receptor alpha(LIFR)	617	214(34.68%)	1(0.48%)

[1] gene of interest [All Fields] AND (tumour[All Fields] OR neoplasms[MeSH Terms] OR neoplasms[All Fields] OR tumor[All Fields] OR cancer[All Fields] OR carcinoma[All Fields]).
[2] gene of interest [All Fields] AND (stomach[All Fields] OR gastric[All Fields]) AND (tumour[All Fields] OR neoplasms[MeSH Terms] OR neoplasms[All Fields] OR tumor[All Fields] OR cancer[All Fields] OR tumor[All Fields] OR carcinoma[All Fields]).
[*] Ratio1 = #(gene of interest-cancer related literatures)/#(gene of interest literatures).
[**] Ratio2 = #(gene of interest-gastric cancer related literatures)/#(gene of interest-cancer related literatures)

TABLE 12
Literature Mining for Predicted Biomarkers for ALL1 (acute lymphocytic leukemia) in PubMed

Probeset ID	Gene symbol	PubMed hits for gene of interest	PubMed hits for gene of interest and ALL[1] (Ratio1[*])
38319_at	CD3d molecule(CD3D)	74	13(17.56%)

[1] gene of interest [All Fields] AND (leukemia[All Fields]).
[*] Ratio1 = #(gene of ALL related literatures)/#(gene of interest literatures).

biomedical research.

ACKNOWLEDGEMENTS

The authors would like to thank the National Natural Science Foundation of China [Grant number 61472158].

REFERENCES

[1] Diao G, Vidyashankar AN. Assessing genome-wide statistical significance for large p small n problems. Genetics. 2013;194(3):7813.
[2] Philip Chen CL, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci. 2014;275:31447.
[3] Dougherty ER. Small sample issues for microarray-based classification. Comp Funct Genomics. 2001;2(1):2834.
[4] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, 2002;30:4147.
[5] Oreski, S., Oreski, D., & Oreski, G. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Systems with Applications, 2012, 39, 1260512617.
[6] Chen, K.-H., Chen, D. C., Liu, Z. Q., Ma. Gene selection for cancer

identification: A decision tree model empowered by particle swarm optimization algorithm. BMC Bioinformatics, 2014;15:49.

[7] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng. 2005;17(4):491502.

[8] N.Zhou, L.Wang, A modified T-test feature selection method and its application on the HapMap genotype data, Genomics Proteomics Bioinformatics 5(3)(2007) 242249.

[9] Liu H, Setiono R. Chi2: Feature Selection and Discretization of Numeric Attributes[C]// International Conference on TOOLS with Artificial Intelligence, 1995. Proceedings. IEEE, 2002:88.

[10] Peng H, Long F, Ding C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8):1226.

[11] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011;334(6062):151824.

[12] Zhou D N, Chen L, Wu D, et al. Maximal Information Coefficient for Feature Selection for Clinical Document Classification[J]. Acta Physico-Chimica Sinica, 2012, volume 28:963-970(8).

[13] Scheidegger C, Sigg L, Behra R. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.[J]. Brain Research, 1999, 501(2):205-14.

[14] Liu Z, Magder LS, Hyslop T, Mao L. Survival associated pathway identification with group lp penalized global AUC maximization. Algorithms Mol Biol. 2010;5:30.

[15] Ruiquan Ge, Manli Zhou, Youxi Luo, Qinghan Meng, Guoqin Mai, Dongli Ma, Guoqing Wang and Fengfeng Zhou. McTwo: a two-step feature selection algorithm based on maximal information Coefficient. BMC Bioinformatics. 2016; 17:142.

[16] Skalak D B. Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms[J]. Machine Learning Proceedings, 1994:293-301.

[17] Deng HT, Runger G. Feature selection via regularized trees. Ieee Ijcnn. 2012.

[18] Chen, K.-H. et al. (2014). Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm. BMC Bioinformatics, 15, 49.

[19] Jin, C., Jin, S. W., & Qin, L. N. Attribute selection method based on a hybrid BPNN and PSO algorithms. Applied Soft Computing, 2012;12: 21472155.

[20] Li, X., Xiao, N., Claramunt, C., & Lin, H. (2011). Initialization strategies to enhancing the performance of genetic algorithms for the p-median problem. Computers & Industrial Engineering, 61, 10241034.

[21] Subhajit Kar, Kaushik Das Sharma, Madhubanti Maitra. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications. 2015;42: 612627.

[22] Moosa J M, Shakur R, Kaykobad M, et al. Gene selection for cancer classification with the help of bees[J]. Bmc Medical Genomics, 2016, 9(Suppl 2):47.

[23] Dashtban M, Balafar M. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts.[J]. Genomics, 2017, 109(2):91-107.

[24] Stjepan Oreski, Goran Oreski. Genetic algorithm-based heuristic for feature selection in credit risk Assessment. Expert Systems with Applications. 2014;41: 20522064.

[25] Martin Jung, Jakob Zscheischler. A guided hybrid genetic algorithm for feature selection with expensive cost functions. International Conference on Computational Science. 2013;18 :2337 2346.

[26] Wang YX, Huang H. Review on statistical methods for gene network reconstruction using expression data. J Theor Biol. 2014;362:5362.

[27] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics.2001;17(6):50919.

[28] Y. Ding, D.Wilkins, Improving the performance of SVM-RFE to select genes inmicroarray data, BMC Bioinforma. 7 (Suppl. 2) (2006) S12.

[29] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification using Support Vector Machines[J]. Machine Learning, 2002, 46(1-3):389-422.

[30] C. Furlanello, M. Serafini, S. Merler, G. Jurman, An accelerated procedure for recursive feature ranking on microarray data, Neural Netw. 16 (2003) 641648.

[31] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng. 2005;17(4):491502.

[32] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS. et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8(1):6874.

[33] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002;1(2):2039.

[34] Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood. 2004;103(7):27718.

[35] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature. 2002; 415(6870):43642.

[36] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JG, Sabet H, Tran T, Yu X et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):50311.

[37] Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 2001;61(7):312430.

[38] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 1999;96(12):674550.

[39] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):5317.

[40] Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy Jr JD. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. N Engl J Med. 2003;349(26):248394.

[41] Wu YH, Grabsch H, Ivanova T, Tan IB, Murray J, Ooi CH, Wright AI, West NP, Hutchins GGA, Wu J, et al. Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. Gut. 2013;62(8):110011.

[42] Wang GS, Hu N, Yang HH, Wang LM, Su H, Wang CY, Clifford R, Dawsey EM, Li JM, Ding T, et al. Comparison of global gene expression of Gastric Cardia and Noncardia cancers from a high-risk population in China. Plos One. 2013; 8(5):e63826.

[43] Levy H, Wang X, Kaldunski M, Jia S, Kramer J, Pavletich SJ, Reske M, Gessel T, Yassai M, Quasney MW, et al. Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. Genes Immun. 2012;13(8):593604.

[44] Krug T, Gabriel JP, Taipa R, Fonseca BV, Domingues-Montanari S, FernandezCadenas I, Manso H, Gouveia LO, Sobral J, Albergaria I, et al. TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. J Cerebr Blood F Met. 2012;32(6):106172.

[45] Limsoon Wong, CS2220:/introduction to Computational Biology Lecture 4: Gene Expression Analysis.

[46] Bhattacharyya, C., Grate, L. R., Rizki, A., Radisky, D., Molina, F. J., Jordan, M. I., et al. (2003). Simultaneous classification and relevant feature identification in high dimensional spaces: Application to molecular profiling data. Signal Processing, 83, 729743.

[47] Chandra, B., & Gupta, M. (2011). Robust approach for estimating probabilistics in Naive-Bayes classifier for gene expression data. Expert Systems with Applications, 38, 12931298

[48] Fu, L. M., & Fu-Liu, C. S. (2005). Evaluation of gene importance in microarray data based upon probability of selection. BMC Bioinformatics, 6, 67.
[49] Ganesh Kumar, P., Aruldoss Albert Victoire, T., Renukadevi, P., & Devaraj, D. (2012). Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. Expert Systems with Applications, 39, 18111821.
[50] Ji, G., Yang, Z., & You, W. (2011). PLS-based gene selection and identification of tumor-specific genes. IEEE Transactions on Systems, Man and Cybernetics Part C:Applications and Reviews, 41(6), 830841.
[51] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expressing profiling and artificial neural network. Nature Medicine, 7, 673679.
[52] Lee, C. P., Lin, W. S., Chen, Y. M., & Kuo, B. J. (2011). Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Systems with Applications, 38, 46614667.
[53] Li, X., & Shu, L. (2009). Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis. Expert Systems with Applications, 36, 76447650.
[54] Mohamad, M. S., Omatu, S., Deris, S., & Yoshioka, M. (2011). A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. IEEE Transactions on Information Technology in Biomedicine, 15(6), 813822.
[55] Pal, N. R., Aguan, K., Sharma, A., & Amari, S. (2007). Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinformatics, 8, 5.
[56] Sharma, A., Imoto, S., & Miyano, S. (2012). A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(3), 754764.
[57] Shen, Q., Shi, W. M., & Kong, W. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Computational Biology and Chemistry, 2008;32:5360
[58] Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS, 99(10), 65676572.
[59] Wong, T. T., & Liu, K. L. (2010). A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection. Expert Systems with Applications, 37, 21442149.
[60] Yang, K., Cai, Z., Li, J., & Lin, G. (2006). A stable gene selection in microarray data analysis. BMC Bioinformatics, 7, 228.
[61] Zainuddin, Z., & Ong, P. (2011). Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. Expert Systems with Applications, 38, 1371113722.
[62] Hossin, M, and M. N. Sulaiman. "A Review on Evaluation Metrics for Data Classification Evaluations." International Journal of Data Mining & Knowledge Management Process 5.2(2015):01-11.
[63] Lin, W. J., and J. J. Chen. "Class-imbalanced classifiers for high-dimensional data." Briefings in Bioinformatics 14.1(2013):13.
[64] Vihinen, Mauno. "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis." Bmc Genomics 13.S4(2012):S2.
[65] Guo P, Luo Y, Mai G, Zhang M, Wang G, Zhao M, Gao L, Li F, Zhou F. Gene expression profile based classification models of psoriasis. Genomics. 2014;103(1):4855.
[66] Duque-Pintor, Francisco, et al. "A New Methodology Based on Imbalanced Classification for Predicting Outliers in Electricity Demand Time Series." Energies 9.9(2016):752.
[67] N.V. Chawla, N. Japkowicz and A. Kolcz, Editorial: Special issue on learning from imbalanced data sets, SIGKDD Explorations, 6 (2004) 1-6.
[68] Q. Gu, L. Zhu and Z. Cai, Evaluation Measures of the Classification Performance of Imbalanced Datasets, in Z. Cai et al. (Eds.) ISICA 2009, CCIS 51. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 461-471.
[69] R. Ranawana, and V. Palade, Optimized precision-A new measure for classifier performance evaluation, in Proc. of the IEEE World Congress on Evolutionary Computation (CEC 2006), 2006, pp. 2254-2261.
[70] Zhang H. Exploring conditions for the optimality of Naive bayes. Int J Pattern Recogn. 2005;19(2):18398.
[71] Prati R C, Batista G E A P A, Monard M C. Data mining with imbalanced class distributions: concepts and methods[C]// Indian International Conference on Artificial Intelligence, Iicai 2009, Tumkur, Karnataka, India, December. DBLP, 2004:359-376.
[72] Goldberg D E, Deb K. A Comparative Analysis of Selection Schemes Used in Genetic Algorithms[J]. Foundations of Genetic Algorithms, 1991, 1:69-93.