Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Gene expression

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

Cheng Peng [1], Xinyu Wu [1], Wen Yuan [1], Xinran Zhang [1], Yu Zhang [1], and Ying Li [1,*]

[1] 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.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Microarray gene expression data has become a topic of great interest for cancer classification. But due to the "large p, small n" paradigm of limited bio-samples and high-dimensional data, feature selection becomes a demanding task which aims at selecting minimal discriminatory genes most associated with the phenotypes. Feature selection is still a challenging problem due to NP time complexity. Most of the existed feature selection algorithms utilize heuristic rules.

Results: A multilayer recursive feature elimination algorithm based on an embedded integer-coded genetic algorithm with dynamic-length chromosome, MGRFE, is proposed, which aims at selecting the gene combination with minimal size and maximal information. Based on 19 benchmark cancer 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 relevances to cancer phenotypes. MGRFE selects a relatively small feature subset with excellent cancer classification performance through comprehensive comparison with mostly state-of-the-art methods on large-scale datasets. MGRFE can be regarded as an promising method for feature selection on high-dimensional dataset especially gene expression data.

Availability: The source codes and all 19 datasets used in this paper are available at http://ml.jlu.edu.cn//MGRFE.

Contact: liying@jlu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics*

1 Introduction

One chief challenge in bioinformatics is the "large p small n" paradigm (Diao and Vidyashankar, 2013), on account of ever-increasing high-dimension data and limited available experimental samples (Chen and Zhang, 2014). It is difficult to acquire sufficient and appropriate bio-samples due to high expense of sample collection and other various factors (Dougherty, 2001). 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 huge feature range is a demanding job (Armstrong *et al.*, 2002). Feature selection is commonly used in pre-processing for high dimensional data analysis. 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 (Oreski *et al.*, 2012), 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 have proposed enormous feature selection algorithms. Without well performed methods of feature selection, several unexpected problems associated with poor scalability and incomplete information appear. The computational complexity of feature selection belongs to NP-hard problems (Chen *et al.*, 2014a). 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 (Liu and Yu, 2005).

Filter algorithms generally evaluate and select attribute indexes according to the inherent characteristics of datasets, and then rank all features so as to form an optimal subset of original features. Up to now, lots of filter algorithms have been designed, such as methods based on T-test (Zhou and Wang, 2007), Chi2 (Liu and Setiono, 1995), mutual information (Peng *et al.*, 2005) and maximal information coefficient (MIC) (Reshef *et al.*, 2011; Lin *et al.*, 2012), Signal-to-Noise-Ratio (Golub *et al.*, 1999a) and so forth. As there is no classification algorithms involved in the filter algorithm, its computational speed is quick. Filter algorithm is suitable for large datasets. Whereas this kind of methods are probable to add redundant features into the chosen feature subsets, which will lead to inaccessibility of the optimal results. In gene expression datasets, genes in a cell interact with others to complete certain biological function. But the filter methods select genes individually rather than select gene combinations, which is another shortcoming (Liu *et al.*, 2010). The top-ranked features by filter algorithm are not always the best feature combinations for classification (Ge *et al.*, 2016).

Wrapper algorithms usually employ classification models and utilize heuristic rules to select feature subsets by the classification performance, 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 (Skalak, 1994), regularized random forest (RRF) (Deng and Runger, 2012), particle swarm optimization (PSO) (Chen *et al.*, 2014b; Jin *et al.*, 2012) and genetic algorithm (GA) (Li *et al.*, 2011). With the rapid development of heuristic rules and evolutionary strategies, various swarm intelligence algorithms have been applied to feature selection. Kar *et al.* (2015) proposed a particle swarm optimization method based on adaptive K-nearest neighborhood (KNN) to identify a minimum meaningful feature subset in gene expression datasets. Moosa *et al.* (2016) presented a modified artificial bee colony algorithm (ABC) to select minimum number of genes with high predictive accuracy for cancer classification. Oreski and Oreski (2014) designed a hybrid genetic algorithm with neural networks to identify an optimum feature subset with high classification accuracy and scalability in credit risk assessment. Jung and Zscheischler (2013) described a guided hybrid GA to minimize the number of cost function evaluations. However, all these feature selection methods based on swarm intelligence algorithms utilize the binary encoding method and lack 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 (Guyon *et al.*, 2002; Ge *et al.*, 2016; Moosa *et al.*, 2016; Dashtban and Balafar, 2017). 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), the slow convergence speed and 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, which can give an explicit recursive reduce to features by removing features with least weights (Ding and Wilkins, 2006; Guyon *et al.*, 2002; Furlanello *et al.*, 2003; Guo *et al.*, 2014).

Hybrid algorithms are the integration of filter and wrapper strategies (Liu and Yu, 2005), 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 the optimal feature selection. Thus, hybrid algorithm can take both advantages of filter and wrapper techniques. our method 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 gene subset with optimal classification performance. Two filter techniques, T-test and MIC, are used in the 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 comprehensively compared our results with 20 representative works using the same datasets. The experiment and comparison results suggest that our method far outweighs most of popular feature selection algorithms in this field, 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 flowchart of whole work in this study is in Fig. 1.

2 Materials and Methods

2.1 Materials

The 19 benchmark cancer gene expression datasets on the problem of gene selection for cancer classification (Ge *et al.*, 2016; Kar *et al.*, 2015) are used to validate our proposed method. The 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 (Ge *et al.*, 2016) and (Kar *et al.*, 2015) and is named as Leuk and ALL_AML respectively but they are the same dataset in actual.

Dataset One consists of the 17 binary classification datasets used in (Ge *et al.*, 2016), which includes DLBCL (Shipp *et al.*, 2002), Prostate (Singh *et al.*, 2002), ALL(divided into four parts based on different phenotypes) (Chiaretti *et al.*, 2004), CNS (Pomeroy *et al.*, 2002) , Lymphoma (Alizadeh *et al.*, 2000), Adenoma (Notterman *et al.*, 2001), Colon (Alon *et al.*, 1999), Leukaemia (Golub *et al.*, 1999b), Myeloma (Tian *et al.*, 2003), Gastric (Wu *et al.*, 2012), Gastric1/Gastric2 (Wang *et al.*, 2013), T1D (Levy *et al.*, 2012) and Stroke (Krug *et al.*, 2012). 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 three typical benchmark datasets used in (Kar *et al.*, 2015) including two multi-class datasets SRBCT (Khan *et al.*, 2001) and MLL (Armstrong *et al.*, 2002) and one binary dataset ALL_AML. The brief description of these three datasets is shown in Table 2. There have been lots of experiments focusing on gene selection for cancer classification performed on these 3 datasets in (Bhattacharyya *et al.*, 2003; Chandra and Gupta, 2011; Fu and Fu-Liu, 2005; Kumar *et al.*, 2012; Ji *et al.*, 2011; Khan *et al.*, 2001; Lee *et al.*, 2011; Li and Shu, 2009; Mohamad *et al.*, 2011; Pal *et al.*, 2007; Sharma *et al.*, 2012; Shen *et al.*, 2008; Tibshirani *et al.*, 2002; Wong and Liu, 2010; Yang *et al.*, 2006;

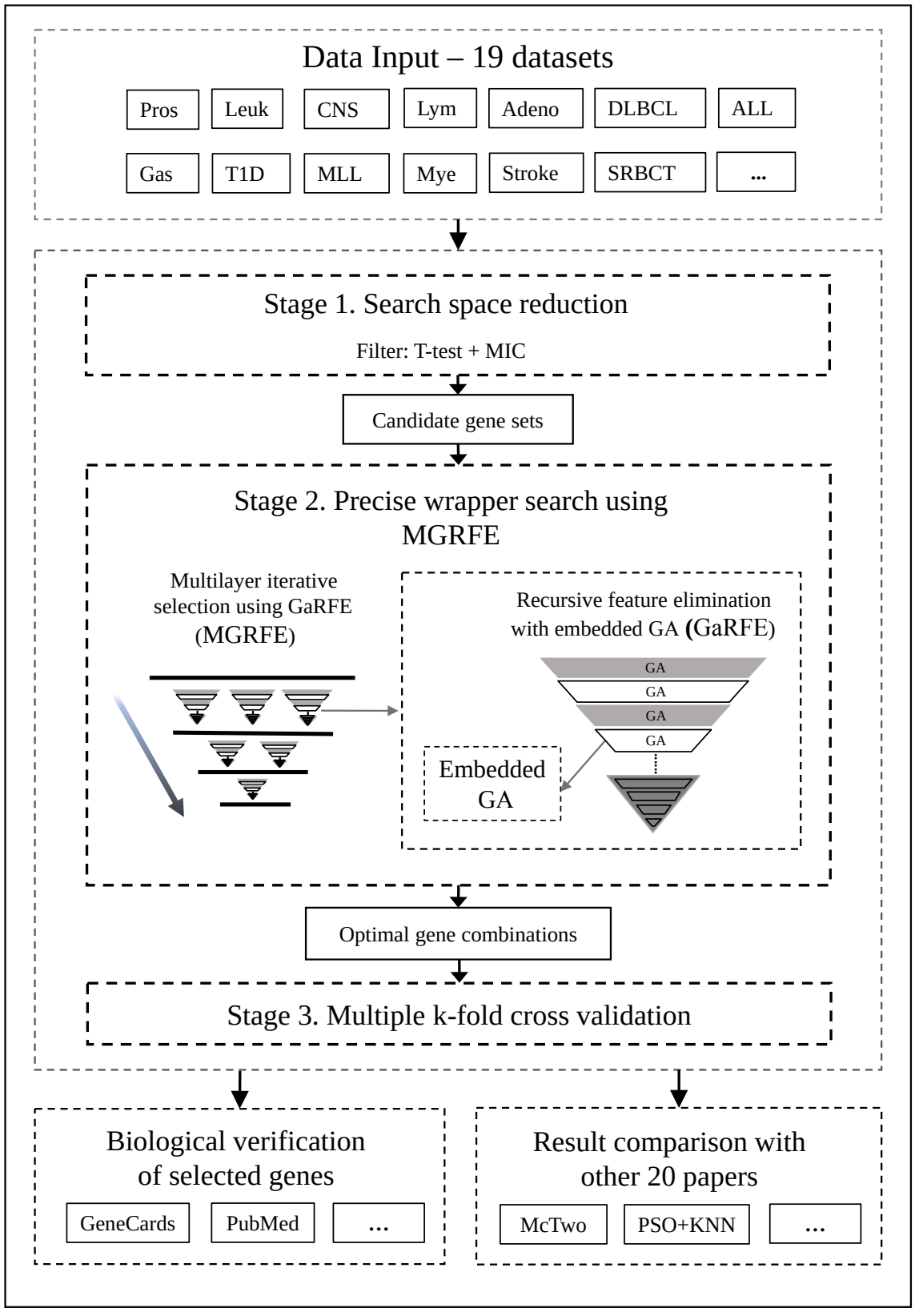

Fig. 1. The flowchart of whole work in this study

Zainuddin and Ong, 2011). We will offer detailed performance comparison between all these methods and the proposed method in later results section.

2.2 Classification performance measurements

According to the previous researches (Hossin and Sulaiman, 2015; Chawla *et al.*, 2004; Gu *et al.*, 2009; Ranawana and Palade, 2006), the accuracy (Acc) is the most widely used evaluation metric for both binary and multi-class classification tasks, thus it is regarded as the main performance metric in comparison between different methods. For the 17 binary datasets including many imbalanced datasets, this study also employed another five wildly used evaluation metrics of Sn, Sp, Avc, MCC and AUC rather than only using Acc to measure the performance of the proposed method and handle data imbalance problem (Hossin and Sulaiman, 2015; Lin and Chen, 2012; Vihinen, 2012; Guo *et al.*, 2014; Duque-Pintor *et al.*, 2016). 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 named Acc and average accuracy of two classes is defined as Avc. The metric Matthews correlation coefficient (MCC) can capture all elements in the confusion matrix other than Acc and is a balanced and comprehension metric. The area under the ROC curve, AUC, can effectively measure the discrimination ability of the classification model (Fawcett, 2006). The calculation formulas of the used metrics are in Equation (1). On the 2 multi-class datasets, for unity and convenience, only accuracy (Acc) is used because the other methods all only calculated this metric.

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

2.3 Method

The proposed method is divided into three stages as shown in Fig. 2 to find minimal discriminatory gene combinations among several thousands of genes in the microarray gene datasets. In the following, the flow of the proposed MGRFE is presented. Moreover, the pseudocode of MGRFE is showed in Algorithm 1.

(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.

Algorithm 1: MGRFE: multilayer iterative feature selection using GaRFE

```
Input  : A microarray gene expression dataset
Output : The optimal gene feature combination for cancer
         classification
T-test based gene ranking to generate the candidate gene set G;
MIC based gene ranking to narrow G;
Set GC, the list of optimal gene combinations in MGRFE, to
empty;
while maximal iterative layer number not reached do
    Initialize and run a layer of GaRFE (Algorithm 2) based on G;
    for each GaRFE do
        Add the returned optimal gene combinations to GC;
    Sort the optimal gene combinations in GC and just preserve the
    top ranking ones;
    Using the genes only in the top ranked gene combinations in
    GC to form a reduced G;
Multiple k-fold CV on the gene combinations in GC;
Return the final selected gene combination;
```

2.3.1 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 the 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 non-linear relations. Firstly, perform T-test on all genes and give them a ascending sort according to their p-values, then the top-ranked significant features with p-values less than 0.05 are preserved. And the upper limit of kept features after the T-test is 1000. That is when there are more than 1000 features with p-values less than 0.05, only the top 1000 with lower p-values are kept. If the preserved features after the T-test screening are less than 500 features, they are all kept directly and definitively to form the candidate gene set without MIC based feature selection, otherwise, the MIC based selection should be followed. 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 higher MIC values. With regards to the 2 multi-class datasets, candidate gene set is obtained according to the descending order of MIC values of all genes.

2.3.2 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 the 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, the inverted triangle in MGRFE as shown in Fig. 1, is the recursive feature elimination process with each stair being embedded genetic algorithm. Embedded GA is the integer-coded genetic algorithm with dynamic-length chromosome. The key of MGRFE is GaRFE at each layer, where embedded GA is responsible for generating the optimal gene combinations and RFE process is responsible for cutting down the gene number. Therefore, our method can find gene combinations with both significantly reduced sizes and excellent classification performances.

Embedded genetic algorithm In our method, the embedded modified GA using variable length integer-coded chromosome is embedded in the recursive feature elimination process as each stair in the inverted triangle of

Table 1. Summary of the 17 binary classification datasets used in Dataset One from (Ge et al., 2016)

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 (Ge *et al.*, 2016).

[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 (Kar et al., 2015)

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 explainations for TP, FP, TN, and 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)

GaRFE. The embedded GA includes the following steps. First, initialize the GA population which are a certain amount of 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 represent best gene combinations to GaRFE. The pseudocode of the embedded GA is presented in Algorithm 3.

To embed GA in the RFE process and achieve the goal of finding minimal informative genes, we made some modifications on the original GA. The embedded GA uses variable length integer encoding method for the chromosome in individual 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 adjacent GA runs 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 be adjusted to adapt to the variable length integer encoding method. One main problem to be concerned is avoiding duplicated genes in every

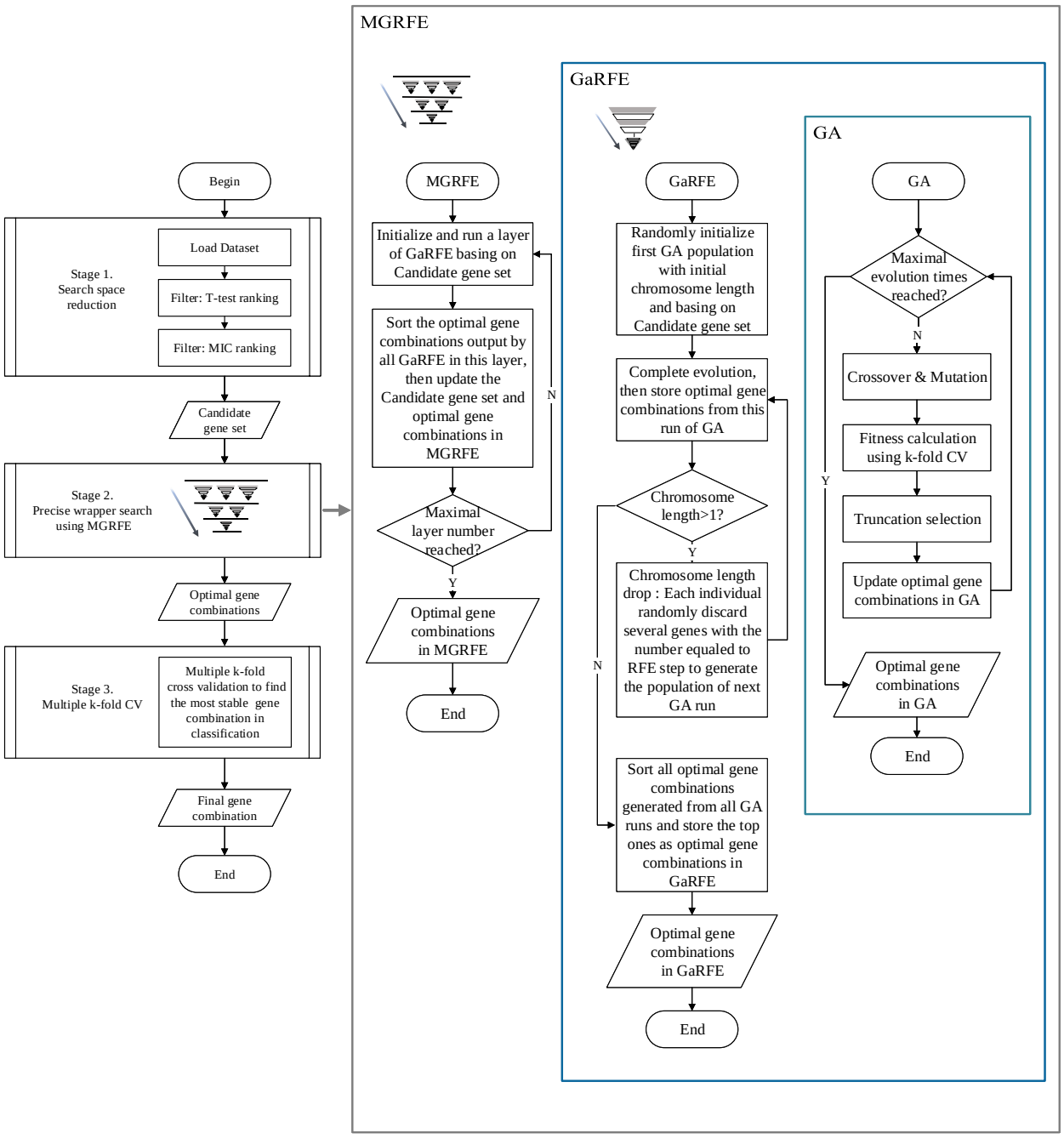

Fig. 2. The flowchart of MGRFE which is divided into 3 stages: search space reduction, precise wrapper search using MGRFE and multiple k-fold CV. The stage 2 is the core of MGRFE, which includes two key processes: GaRFE and embedded modified GA.

individual, which leads 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 uses single point crossover, 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 Equation (2) Acc represents the average accuracy from 5-fold cross validation, where α is an adjustment coefficient to deal 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 (Prati *et al.*, 2009) for $Avc = (Sn + Sp)/2$ takes the correct prediction proportion of both two sample classes into consideration. The

Algorithm 2: GaRFE: recursive feature elimination with embedded GA

Input : Candidate gene set G, Maximal chromosome length L
Output: The optimal gene feature combinations in GaRFE
Randomly generate the fisrt GA population P from G with chromosome length equal to L;
Set GC, the list of optimal gene combinations in MgRFE, to empty;
do
  Excute embedded GA (Algorithm 3) using the population P;
  Add the returned gene combinations by GA to GC;
  if *current chromosome length* > 1 **then**
    Chromosome length drop: each individual in P randomly discard several genes with the number equaled to RFE step;
while *the chromosome length of current GA population* > 1 ;
Sort the optimal gene combinations in GC and just preserve the top ranking ones;
Return the optimal gene combinations in GC;

Algorithm 3: Embedded GA

Input : GA population P, Maximal evolution times T
Output: Updated P, The optimal gene feature combinations in GA
Set GC, the list of optimal gene combinations in GA, to empty;
while *maximal evolution times* T *not reached* **do**
  Perform mutation operator;
  Perform crossover operator;
  Fitness calculation of individuals by k-fold CV;
  Truncation selection to form the updated P;
Sort the individuals in P and select top ones to form GC;
Return the updated population P and optimal gene combinations in GC;

fitness definition can be regarded as a new way to help tackle the data imbalanced problem. For balanced datasets, F can be directly calculated as Acc. 5-fold cross validation is used to calculate the fitness value and the employed classifier is Naive Bayes classifier (NBayes) (Zhang, 2005). The sort 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 superior. For two individuals having same fitness values, the one with smaller gene number is superior. Individuals are exactly gene combinations. MGRFE and GaRFE use this sorting rule to rank different gene combinations.

Selection process in embedded GA uses truncation selection method (Blickle and Thiele, 1995), which simply ranks all individuals according to their related fitness values and selects the top individuals to form the next generation. The stopping criteria is evolution time which are set from 1 to 3. In each evolution, perform mutation, crossover, fitness calculation and selection and then update the fittest individuals which are the optimal gene combinations. When the maximal evolution time is reached, embedded GA returns its overall optimal gene combinations to GaRFE.

Recursive feature elimination with embedded modified GA GaRFE is designed as an explicit recursive feature elimination of gene size with embedded modified GA to find minimal discriminatory gene combinations. The pseudocode of GaRFE is presented in Algorithm 2. First, randomly generate the initial GA population based 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 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 between two GA runs, the RFE step, 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 iterative 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 are sorted and the top-ranked combinations are taken as the update of the 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 for further validation.

2.3.3 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 minimal 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

This section presents the performance comparison with the most popular feature selection methods on 19 benchmark datasets. The datasets are divided into Dataset One and Dataset Two. Datasets. In addition, the biological verification of the selected genes is also discussed.

3.1 Results on Dataset One

The results of MGRFE on 17 binary datasets are listed in Table 4 including six evaluation metrics calculated by 5-fold CV and T-test based gene rankings. On total 17 datasets, Acc values are all above 0.9 within 10 genes. Moreover, on 8 of 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 for imbalanced datasets including DLBCL, Colon, Leuk, ALL1, ALL4 and CNS, where Sn, Sp, Avc, MCC and AUC values are relatively high and haven't been influenced by the imbalanced samples. From the T-test based gene rankings, the best gene or feature subset is not always the top-ranked features in T-test, so only filter algorithm can't generate the best feature combination. On 5 of 17 datasets, the top one gene from T-test which has the position numbered 0 appeared in the final selected 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.

3.2 Comparison with other methods on Dataset One

McTwo (Ge *et al.*, 2016) tested all datasets in Dataset One and showed satisfied results. Therefore, we offer the performance comparison between McTwo and MGRFE. Table 5 shows the overall maximal Acc and numbers of selected genes on total 17 datasets by MGRFE and McTwo. For more intuitively comparison, Fig. 4 offers the line chart of maximal Acc achieved by McTwo and MGRFE on 17 datasets, where MGRFE obviously outperforms McTwo with a higher Acc line.

Table 4. Results of MGRFE on 17 datasets in Dataset One

Dataset	Pos/Neg	Genes	Sn	Sp	Acc	Avc	MCC	AUC	T-test based gene rankings
DLBCL	58/19	**3**	1.0	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	0.98	[0, 14, 73, 693]
Colon	40/22	**6**	1.0	0.960	**0.985**	0.980	0.969	0.97	[14, 57, 175, 224, 239, 494]
Leuk	47/25	**2**	1.0	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	0.95	[2, 14, 82, 142, 377, 403, 568]
ALL1	95/33	**1**	1.0	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	0.94	[0, 51, 77, 79, 521, 686, 736, 759]
ALL3	24/101	**8**	0.830	0.950	**0.927**	0.890	0.785	0.93	[3, 51, 74, 141, 487, 509, 714, 769]
ALL4	26/67	**6**	1.0	0.986	**0.990**	0.993	0.978	0.99	[0, 5, 38, 281, 534, 753]
CNS	39/21	**7**	1.0	1.0	**1.0**	1.0	1.0	0.98	[8, 52, 129, 130, 271, 272, 519]
Lym	22/23	**3**	1.0	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	1.0	[467]
Gas	29/36	**3**	1.0	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	0.99	[131, 247, 716]
Gas2	62/62	**2**	1.0	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	0.94	[13, 24, 112, 558, 577, 679, 977]
Stroke	20/20	**4**	1.0	1.0	**1.0**	1.0	1.0	1.0	[0, 22, 128, 275]

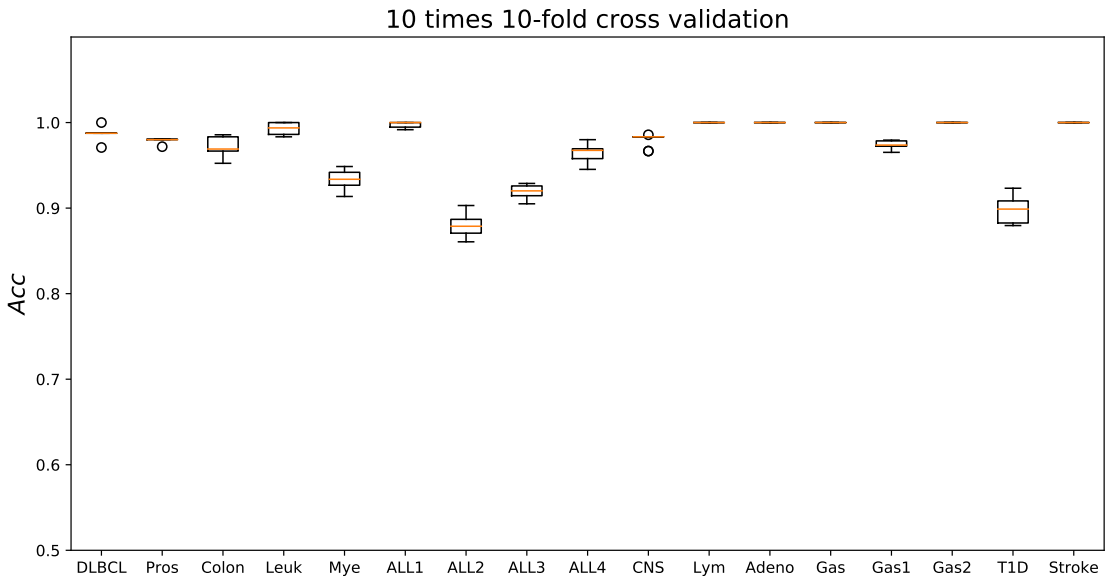

Fig. 3. The distribution of Acc values for 10 times 10-fold cross validation on the selected gene combinations of 17 datasets

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 Acc values by these two algorithms when gene number 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. However, the Acc values by using these gene numbers fall behind our optimal Acc values on these datasets by a large margin, thus MGRFE selected little more genes to achieve the optimal results.

3.3 Results on Dataset Two

The results of MGRFE on 3 typical benchmark datasets including two multiclass datasets is presented here. 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 processes at the first layer of MGRFE on these datasets, from which the Acc values of the best GA individuals in GaRFE are kept 1.0 on the marjority of gene numbers and only begin to drop when the gene number is significantly reduced. 10 times 10-fold cross validation are carried to further validate the final selected gene combinations on Dataset Two as shown in Fig. 6. The Acc by multiple k-fold CV on SRBCT, ALL_AML and MLL are 98.8%, 98.3%, and 99.7% respectively, which shows that MGRFE has high classification stability.

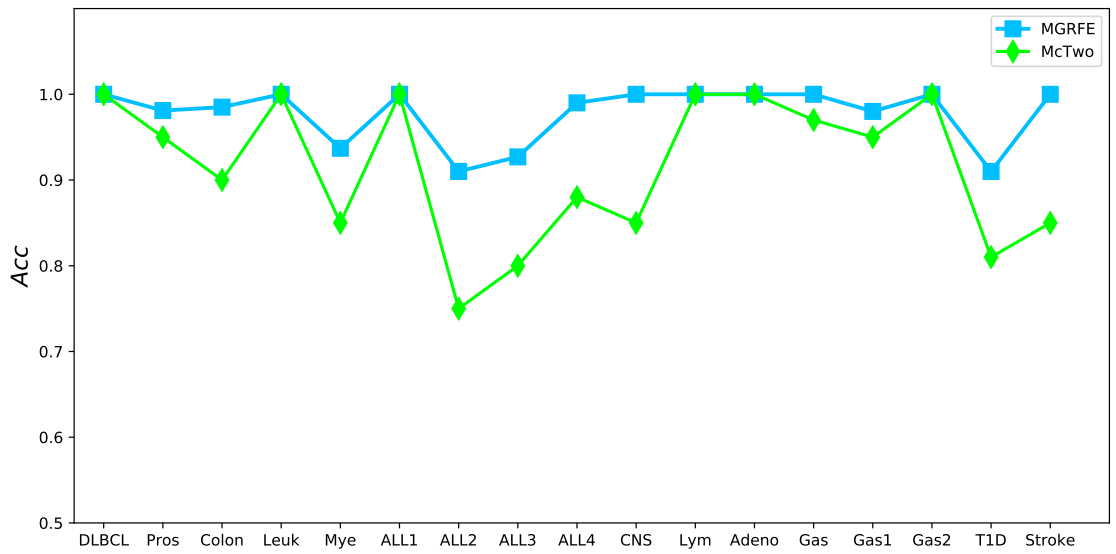

Fig. 4. The line plots of overall maximal accuracies for MGRFE and McTwo on 17 datasets in Dataset One

Table 5. Performance comparison between McTwo and MGRFE on 17 datasets in Dataset One

	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

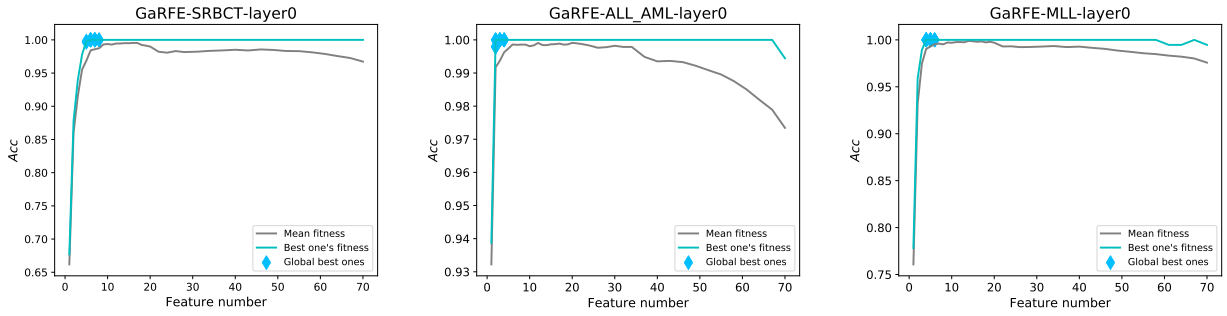

Fig. 5. Three GaRFE processes on the 3 benchmark datasets in Dataset Two

3.4 Comparison with other methods on Dataset Two

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

For SRBCT dataset, Khan *et al.* (2001) applied artificial neural network (ANN) and selected 96 genes to achieve 100% Acc; Tibshirani *et al.* (2002) used the nearest shrunken centroid based method and achieved 100% Acc by 43 genes; Fu and Fu-Liu (2005) used SVM-RFE and achieved 100% Acc by 19 genes; Pal *et al.* (2007) applied multi-layered perceptron and non-Euclidean relational fuzzy c-means clustering and find 7 genes important for 100% Acc; Mohamad *et al.* (2011) used improved binary PSO and 6 genes were selected; Kar *et al.* (2015) applied PSO and KNN to select 6 genes; Moosa *et al.* (2016) achieved 100% Acc with modified artificial bee colony algorithm by 5 genes; Sharma *et al.* (2012) 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, so MGRFE didn't choose them. For SRBCT dataset, MGRFE selected 5 genes and achieved 100% 5-fold Acc and 98.5% 10 times 10-fold CV Acc.

For ALL_AML dataset, Fu and Fu-Liu (2005) achieved 100% train Acc by 19 genes based on SVM-RFE; Yang *et al.* (2006) applied gene scoring technique and SVM to select 4 genes with 98.6% Acc in leave one out cross validation (LOOCV); Mohamad *et al.* (2011) selected 2 genes to reach 100% CV Acc based on improved binary PSO; Dashtban and Balafar (2017) applied integer encoding GA and SVM to select 15 genes with 100% Acc; Ge *et al.* (2016) designed a MIC based two step method

Table 6. Performance comparison on 5 datasets between MGRFE and McTwo when MGRFE uses the same gene numbers as McTwo used

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

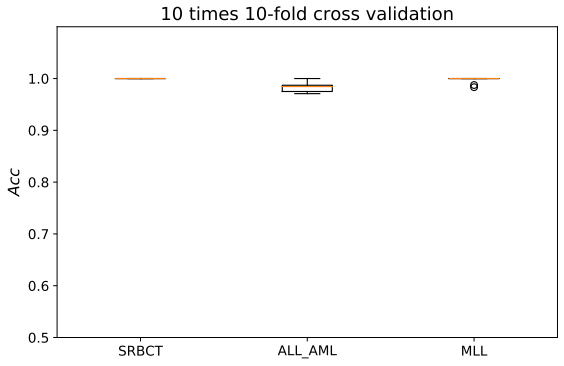

Fig. 6. The distribution of Acc values for 10 times 10-fold cross validation on the selected gene combinations of 3 benchmark datasets in Dataset Two

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.

For MLL dataset, Sharma *et al.* (2012) selected 4 genes with 100% train and test Acc based on SFS, LDA and NCC; Mohamad *et al.* (2011) selected 4 genes with 100% CV Acc based on improved binary PSO; Dashtban and Balafar (2017) applied integer encoding GA and SVM to select 15 genes with 100% Acc; Kar *et al.* (2015) 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.

3.5 Biological inferences of genes selected by MGRFE

The selected genes by MGRFE have closely biological relevances to the phenotypes in gene expression datasets by text mining. In Table 10, 11 and 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 100% 5-fold CV Acc.

For 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, the E protein encoded by TCF3 plays a critical role in lymphopoiesis and is necessary for B and T lymphocyte. TCF3 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). For Gastric dataset, 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 candidates for gastric cancer biological researchers. For 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 (Wong, 2012), it has also been pointed out that the gene CD3D is one ideally discriminatory feature and can be the diagnostic marker when the expression of CD3D is below certain limit. CD3D is involved in T-cell development and signal transduction and defects in this gene will lead to severe combined immunodeficiency.

4 Discussion

In this paper MGRFE, a novel multilayer recursive feature elimination algorithm based on an embedded integer-coded genetic algorithm with dynamic-length chromosome is proposed, which aims at selecting minimal discriminatory genes or features associated closely with the phenotypes. MGRFE provides a complementary feature selection algorithm for high-dimensional data expecially gene expression data analysis and be applied for cancer diagnosis and further biomedical research.

The innovation of our method is to combine the advantages of evolution calculation of the embedded GA and explicit feature decline of the RFE process as GaRFE, which is the feature search unit at each layer of MGRFE. MGRFE can give an explicit feature number decline along with the evolution optimization search and achieve quick convergence speed. Genetic algorithms are probabilistic search algorithms inspired by the natural evolution which include genetic operators of mutation, crossover and selection. Binary encoding is the most common used solution encoding method for GA and other evolution algorithms (Goldberg, 2006). Furthermore, Lee and Antonsson (2000) and Kim and De Weck (2005) presented some kinds of variable length encoding methods for GA, and in the GA population where the length of chromosome varies among different individuals to strengthen the represent ability of chromosome for specific problem's solution. Recently, some evolution based methods have been designed for feature selection, where binary encoding has been extensively considered for solution encoding in (Yang and Honavar, 1998; Jung and Zscheischler, 2013; Oreski and Oreski, 2014; Kar *et al.*, 2015; Moosa *et al.*, 2016). But binary encoding has the shortcomings of the probable existing of irrelevant features in selected feature subset, high time cost for decreasing the feature number and slow convergence speed especially when dealing with high-dimensional data. Instead, our proposed method utilizes variable length integer encoding method in GA and cuts down the encoding length recursively in search process, which could quickly remove the irrelevant and redundant features and converge to the minimal informative feature combination. Moreover, our designed method calls GA many times and in each run of GA, all individuals in the population have the fixed length chromosomes and chromosome length only changes between two adjacent GA runs. A similar work by (Dashtban and Balafar, 2017) also provided an integer-coded genetic algorithm with dynamic-length chromosome and intelligent parameter settings for gene selection on microarray data, but their method lacks the recursive feature decline operation. We have made performance comparison between Dashtban' method and our method in Table 7 and 8, where the results by Dashtban' method on datasets SRBCT and ALL_AML are inferior than the results obtained by our proposed method. Compared to Dashtban' method, our proposed MGRFE can find much smaller gene subset. In addition, it is also worth mentioning that the RFE process in MGRFE doesn't rank genes to remove the least weighted ones as described in (Guyon *et al.*, 2002), but introduces the random strategy to randomly discards a same number of genes in each individual's

Table 7. Performance comparison of the methods on SRBCT dataset.

Experiments	Methods	Genes				CV Acc(%)				Train Acc(%)	Test Acc(%)
Khan *et al.* (2001)	ANN	96				-				100	100
Tibshirani *et al.* (2002)	NSC	43				-				100	100
Fu and Fu-Liu (2005)	SVM-RFE	19				-				100	100
		5CV		LOOCV		5CV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang *et al.* (2006)	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)	FSMLP+NERFCM	7				-				100	100
Ji *et al.* (2011)	PLSVIP	24				-				100	100
	PLSVEG	15				-				100	100
Mohamad *et al.* (2011)	IBPSO	6				100				-	-
Sharma *et al.* (2012)	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	90
	SFS+NNC	4				-				100	95
Zainuddin and Ong (2011)	MSFCM+WNN	10				10CV					
						100				-	-
Li and Shu (2009)	KLLE+LLE+PCA	20				-				100	100
Lee *et al.* (2011)	AGA+KNN	14				-				100	100
Chen *et al.* (2014a)	PSODT	-				5CV					
						92.94				-	-
Kar *et al.* (2015)	PSO+KNN	6				98.0159				100	100
Moosa *et al.* (2016)	ABC	5								100	100
Dashtban and Balafar (2017)	GA+SVM	18								100	100
This paper	**MGRFE**	**5**				**98.8**				**100**	**100**

Table 8. Performance comparison of the methods on ALL_AML (Leukemia) dataset.

Experiments	Methods	Genes				CV Acc(%)				Train Acc(%)	Test Acc(%)
Fu and Fu-Liu (2005)	SVM-RFE	19				-				100	97.06
		5CV		LOOCV		5CV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang *et al.* (2006)	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)	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)	PLSVIP	9				-				100	100
	PLSVEG	8				-				100	100
Mohamad *et al.* (2011)	IBPSO	2				100				-	-
Zainuddin and Ong (2011)	MSFCM+WNN	10				10CV					
						98.61				-	-
Wong and Liu (2010)	Probabilistic mechanism	-				SVM	KN				
						97.38	98.21			-	-
Chandra and Gupta (2011)	RNBC	-				10CV					
						RNBC	NBC	KNN			
						94.29	84.29	85.71		-	-
Kumar *et al.* (2012)	GSA	10				100				-	-
Kar *et al.* (2015)	PSO+KNN	3				95.8868				100	97.0588
Ge *et al.* (2016)	McTwo	2								100	100
Dashtban and Balafar (2017)	GA+SVM	15								100	100
This paper	**MGRFE**	**2**				**98.3**				**100**	**100**

chromosome between two GA runs which proves to be very effective for the embedded GA by our comprehensive experiments on 19 microarray datasets. As for the selection operator of our embedded GA, the frequently used roulette wheel selection (Blickle and Thiele, 1995) is less efficient than the simple truncation selection currently used in MGRFE based on our extra experiments. 1), the differences between the fitness values of all individuals are just slight in most cases. 2), the gap between fitness values only account for few portion in the whole fitness value. These two points provide all individuals with nearly same area occupations in the roulette wheel and lead to the inefficiency of roulette wheel selection. About the parameter setting of MGRFE to make it competent for finding the minimal informative feature subset with fast convergence speed, 1) the evolution times of embedded GA dynamically set just from 1 to 3 (smaller time for larger chromosome length to save time) is enough; 2) the declined feature number, RFE step, between two GA runs dynamically set from 1 to 3 (larger step for larger chromosome length to save time) is qualified; 3) the iterative layer number of MGRFE set 3 with 3/2/1 GaRFE processes in each layer is enough.

The proposed method could be applied to the feature selection and classification problems of high-dimensional data especially the gene

Table 9. Performance comparison of the methods on MLL dataset.

Experiments	Methods	Genes				CV Acc(%)				Train Acc(%)	Test Acc(%)
		5CV		LOOCV		5CV		LOOCV			
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang *et al.* (2006)	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)	SFS+LDA with NCC	4				-				100	100
	SFS+Bayes classifier	4				-				100	100
	SFS+NNC	4				-				100	93
Mohamad *et al.* (2011)	IBPSO	4				100				-	-
Chandra and Gupta (2011)	RNBC	-				10CV					
						RNBC	NBC	KNN			
						87.14	80	68.57		-	-
Chen *et al.* (2014a)	PSODT	âĽ“				5CV					
						100				-	-
Kar *et al.* (2015)	PSO+KNN	4				92.5439				100	100
This paper	**MGRFE**	**3**				**99.7**				**100**	**100**

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).

expression data. For the analysis of high-dimensional data, feature selection is essential, which aims at removing the irrelevant and redundant features, cutting down the dimensionality, and improving the predictive performance and model interpretability. But due to its NP time complexity, feature selection is still a challenging and extensively studied problem in the machine learning and data mining fields. As for the field of bioinformatics, there are numerous of high dimensional biological data in issues like sequence analysis, microarray analysis and spectral analyses which makes feature selection more important and challenging (Saeys *et al.*, 2007). Moreover, the identification of a small subset of informative predictors is deeply desired in the analysis of high-throughput genomic, proteomic and metabolomic data which is characterized by the "large p small n" paradigm (Lin and Chen, 2012). The proposed method could be quite qualified for the feature selection tasks in above high-dimensional datasets due to its excellent performance.

The 19 benchmark microarray datasets for feature selection including multi-class and imbalanced datasets are 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 reach 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 show the robustness for multi-class datasets and imbalanced datasets according to Sn, Sp, Avc, MCC and AUC metrics. Based on classification performance comparison with other 20 methods on the two large Datasets, our proposed method MGRFE is proved to be 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 taken as novel biomarker candidates, which are significant for further biological and medical research. Moreover, MGRFE can contribute to developing 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.

Acknowledgements

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

References

Alizadeh, A. A., *et al.* (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503.

Alon, U., *et al.* (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, **96**(12), 6745–6750.

Armstrong, S. A., *et al.* (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, **30**(1), 41.

Bhattacharyya, C., *et al.* (2003). Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data. *Signal Processing*, **83**(4), 729–743.

Blickle, T. and Thiele, L. (1995). A comparison of selection schemes used in genetic algorithms.

Chandra, B. and Gupta, M. (2011). Robust approach for estimating probabilities in naïve–bayes classifier for gene expression data. *Expert Systems with Applications*, **38**(3), 1293–1298.

Chawla, N. V., *et al.* (2004). Editorial:special issue on learning from imbalanced data sets. *Acm Sigkdd Explorations Newsletter*, **6**(1), 1–6.

Chen, C. P. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, **275**, 314–347.

Chen, K.-H., *et al.* (2014a). Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC bioinformatics*, **15**(1), 49.

Chen, K.-H., *et al.* (2014b). Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC bioinformatics*, **15**(1), 49.

Chiaretti, S., *et al.* (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**(7), 2771–2778.

Dashtban, M. and Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics*, **109**(2), 91–107.

Deng, H. and Runger, G. (2012). Feature selection via regularized trees. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE.

Diao, G. and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large p small n problems. *Genetics*, **194**(3), 781–783.

Ding, Y. and Wilkins, D. (2006). Improving the performance of svm-rfe to select genes in microarray data. *BMC bioinformatics*, **7**(2), S12.

Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, **2**(1), 28–34.

Duque-Pintor, F. J., *et al.* (2016). A new methodology based on imbalanced classification for predicting outliers in electricity demand time series. *Energies*, **9**(9), 752.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, **27**(8), 861–874.

Fu, L. M. and Fu-Liu, C. S. (2005). Evaluation of gene importance in microarray data based upon probability of selection. *BMC bioinformatics*, **6**(1), 67.

Furlanello, C., *et al.* (2003). An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks*, **16**(5), 641–648.

Ge, R., *et al.* (2016). Mctwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC bioinformatics*, **17**(1), 142.

Goldberg, D. E. (2006). Genetic algorithms in search, optimization and machine learning. **xiii**(7), 2104âŁ“2116.

Golub, T. R., *et al.* (1999a). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286**(5439), 531–537.

Golub, T. R., *et al.* (1999b). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, **286**(5439), 531–537.

Gu, Q., *et al.* (2009). Evaluation measures of the classification performance of imbalanced data sets. *Computational intelligence and intelligent systems*, pages 461–471.

Guo, P., *et al.* (2014). Gene expression profile based classification models of psoriasis. *Genomics*, **103**(1), 48–55.

Guyon, I., *et al.* (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, **46**(1), 389–422.

Hossin, M. and Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, **5**(2), 1.

Ji, G., *et al.* (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), 830–841.

Jin, C., *et al.* (2012). Attribute selection method based on a hybrid bpnn and pso algorithms. *Applied Soft Computing*, **12**(8), 2147–2155.

Jung, M. and Zscheischler, J. (2013). A guided hybrid genetic algorithm for feature selection with expensive cost functions. *Procedia Computer Science*, **18**, 2337–2346.

Kar, S., *et al.* (2015). Gene selection from microarray gene expression data for classification of cancer subgroups employing pso and adaptive k-nearest neighborhood technique. *Expert Systems with Applications*, **42**(1), 612–627.

Khan, J., *et al.* (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, **7**(6), 673.

Kim, I. Y. and De Weck, O. (2005). Variable chromosome length genetic algorithm for progressive refinement in topology optimization. *Structural and Multidisciplinary Optimization*, **29**(6), 445–456.

Krug, T., *et al.* (2012). Ttc7b emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. *Journal of Cerebral Blood Flow & Metabolism*, **32**(6), 1061–1072.

Kumar, P. G., *et al.* (2012). Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Systems with Applications*, **39**(2), 1811–1821.

Lee, C. and Antonsson, E. (2000). Variable length genomes for evolutionary algorithms. In *GECCO*, volume 2000, page 806.

Lee, C.-P., *et al.* (2011). Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. *Expert Systems with Applications*, **38**(5), 4661–4667.

Levy, H., *et al.* (2012). Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. *Genes and immunity*, **13**(8), 593.

Li, X. and Shu, L. (2009). Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis. *Expert Systems with Applications*, **36**(4), 7644–7650.

Li, X., *et al.* (2011). Initialization strategies to enhancing the performance of genetic algorithms for the p-median problem. *Computers & Industrial Engineering*, **61**(4), 1024–1034.

Lin, C., *et al.* (2012). Maximal information coefficient for feature selection for clinical document classification. In *ICML Workshop on Machine Learning for Clinical Data. Edingburgh, UK*.

Lin, W.-J. and Chen, J. J. (2012). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics*, **14**(1), 13–26.

Liu, H. and Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *Tools with artificial intelligence, 1995. proceedings., seventh international conference on*, pages 388–391. IEEE.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, **17**(4), 491–502.

Liu, Z., *et al.* (2010). Survival associated pathway identification with group l p penalized global auc maximization. *Algorithms for Molecular Biology*, **5**(1), 30.

Mohamad, M. S., *et al.* (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), 813–822.

Moosa, J. M., *et al.* (2016). Gene selection for cancer classification with the help of bees. *BMC medical genomics*, **9**(2), 47.

Notterman, D. A., *et al.* (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer research*, **61**(7), 3124–3130.

Oreski, S. and Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert systems with applications*, **41**(4), 2052–2064.

Oreski, S., *et al.* (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with applications*, **39**(16), 12605–12617.

Pal, N. R., *et al.* (2007). Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. *BMC bioinformatics*, **8**(1), 5.

Peng, H., *et al.* (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, **27**(8), 1226–1238.

Pomeroy, S. L., *et al.* (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**(6870), 436–442.

Prati, R. C., *et al.* (2009). Data mining with imbalanced class distributions: concepts and methods. In *IICAI*, pages 359–376.

Ranawana, R. and Palade, V. (2006). Optimized precision-a new measure for classifier performance evaluation. In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, pages 2254–2261. IEEE.

Reshef, D. N., *et al.* (2011). Detecting novel associations in large data sets. *science*, **334**(6062), 1518–1524.

Saeys, Y., *et al.* (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, **23**(19), 2507–2517.

Sharma, A., *et al.* (2012). A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **9**(3), 754–764.

Shen, Q., *et al.* (2008). Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. *Computational Biology and Chemistry*, **32**(1), 53–60.

Shipp, M. A., *et al.* (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, **8**(1), 68–74.

Singh, D., *et al.* (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, **1**(2), 203–209.

Skalak, D. B. (1994). Prototype and feature selection by sampling and random mutation hill climbing algorithms. In *Proceedings of the eleventh international conference on machine learning*, pages 293–301.

Tian, E., *et al.* (2003). The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine*, **349**(26), 2483–2494.

Tibshirani, R., *et al.* (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, **99**(10), 6567–6572.

Vihinen, M. (2012). How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis. *BMC genomics*, **13**(4), S2.

Wang, G., *et al.* (2013). Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. *PloS one*, **8**(5), e63826.

Wong, L. (2012). Lecture 4: Gene expression analysis.

Wong, T.-T. and Liu, K.-L. (2010). A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection. *Expert Systems with Applications*, **37**(3), 2144–2149.

Wu, Y., *et al.* (2012). Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut*, pages gutjnl–2011.

Yang, J. and Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems and their Applications*, **13**(2), 44–49.

Yang, K., *et al.* (2006). A stable gene selection in microarray data analysis. *BMC bioinformatics*, **7**(1), 228.

Zainuddin, Z. and Ong, P. (2011). Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. *Expert Systems with Applications*, **38**(11), 13711–13722.

Zhang, H. (2005). Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, **19**(02), 183–198.

Zhou, N. and Wang, L. (2007). A modified t-test feature selection method and its application on the hapmap genotype data. *Genomics, proteomics & bioinformatics*, **5**(3), 242–249.