# Supplementary For MGRFE: multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification

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

## S1. The datasets used in this study

This study used total 19 benchmark microarrays subdivided into two large Datasets to validate the performance of our proposed MGRFE. In Tables 1 and 2, we provide the brief description of each dataset in Dataset One and Dataset Two.

## S2. Pseudocode of the proposed MGRFE

In the main manuscript, we provided the flow chart of the proposed MGRFE in Fig. 2. Here, we supplement the pseudocodes of our methodology. Pseudocode 1 describes the complete procedure of MGRFE. Pseudocodes 2 and 3 explain the two key processes of MGRFE: GA-RFE and embedded GA.

**Pseudocode 1:** MGRFE: multilayer iterative feature selection using GA-RFE

```
Input  : A microarray gene expression data
Output: The optimal gene feature combination
        for phenotype classification
The t-test-based gene ranking to generate the
 candidate gene set G;
The MIC-based gene ranking to narrow G;
Set GC, the list of optimal gene combinations
 in MGRFE, to empty;
while the maximal iterative layer number not
 reached do
    Initialize and run a layer of GA-RFE
     (Pseudocode 2) based on G;
    for each GA-RFE do
        Add the returned optimal gene
         combinations to GC;
    Sort the optimal gene combinations in GC
     and only preserve the top ranked ones;
    Use the genes 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;
```

## S3. Implementation notes and Computation time

We implemented the proposed MGRFE in Python version 3.6.0 environment (https://www.python.org/) on a common laptop computer with Intel(R) Core(TM) i5-4210U CPU and 8G memory. The Python SciPy package version 0.19.0 [3] was involved in the *t*-test process, and minepy package version 1.2.0 [4] was used to perform the MIC calculation. Some parameter settings about MGRFE: 1) the evolution iteration number of embedded GA was dynamically set to 1 to 3 (smaller iteration number used for larger chromosome length to save time); 2) the reduced feature number between two GA runs, the RFE step, was dynamically set to 1 to 3 (larger reduction step used for larger chromosome length to save time); and 3) the iterative layer number of MGRFE being 3 with three, two and


- *C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Zhang, and Y. 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.*
- *Y. Li is the correspondence author. Email: liying@jlu.edu.cn.*

TABLE 1
Summary of the 17 binary classification datasets in Dataset One from ref. [1]

| ID | Dataset | Samples | Features | Summary |
|---|---|---|---|---|
| 1 | DLBCL[1] | 77 | 7 129 | DLBCL patients (58) and follicular lymphoma (19) |
| 2 | Pros(Prostate)[1] | 102 | 12 625 | prostate (52) and non-prostate (50) |
| 3 | Colon[2] | 62 | 2 000 | tumour (40) and normal (22) |
| 4 | Leuk(Leukaemia)[2] | 72 | 7 129 | ALL (47) and AML (25) |
| 5 | Mye(Myeloma)[3] | 173 | 12 625 | presence (137) and absence (36) of focallesions of bone |
| 6 | ALL1[1] | 128 | 12 625 | B-cell (95) and T-cell (33) |
| 7 | ALL2[1] | 100 | 12 625 | patients that did (65) and did not (35) relapse |
| 8 | ALL3[1] | 125 | 12 625 | with (24) and without (101) multidrug resistance |
| 9 | ALL4[1] | 93 | 12 625 | with (26) and without (67) the t(9;22) chromosome translocation |
| 10 | CNS[1] | 60 | 7 129 | medulloblastoma survivors (39) and treatment failures (21) |
| 11 | Lym(Lymphoma)[1] | 45 | 4 026 | germinalcentre (22) and activated B-like DLBCL (23) |
| 12 | Adeno(Adenoma)[1] | 36 | 7 457 | colon adenocarcinoma (18) and normal (18) |
| 13 | Gas(Gastric)[3] | 65 | 22 645 | tumors (29) and non-malignants (36) |
| 14 | Gas1(Gastric1)[3] | 144 | 22 283 | non-cardia (72) of gastric and normal (72) |
| 15 | Gas2(Gastric2)[3] | 124 | 22 283 | cardia (62) of gastric and normal (62) |
| 16 | T1D[3] | 101 | 54 675 | T1D (57) and healthy control (44) |
| 17 | Stroke[3] | 40 | 54 675 | ischemic stroke (20) and control (20) |

In Tables 1 and 2, "Samples" and "Features" indicate the total sample number and feature number of each dataset, and "Summary" column describes the sample classes and the related sample numbers in parenthesis.
[1] These datasets were retrieved from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.
[2] Colon and Leuk datasets were downloaded from the R/Bioconductor packages *colonCA* and *golubEsets*, respectively.
[3] These datasets were downloaded from https://www.ncbi.nlm.nih.gov/geo/.

TABLE 2
Summary of the 3 classification datasets in Dataset Two from ref. [2]

| ID | Dataset | Classes | Samples | Features | Summary |
|---|---|---|---|---|---|
| 1 | SRBCT[1] | 4 | 88 | 2 308 | EWS (29), NHL (11), NB (18) and RMS (25) |
| 2 | ALL_AML[2] | 2 | 72 | 7 129 | ALL (47) and AML (25) |
| 3 | MLL[3] | 3 | 72 | 12 582 | ALL (24), MLL (20) and AML (28) |

[1] SRBCT dataset was downloaded from http://research.nhgri.nih.gov/microarray/Supplement/. This dataset includes 88 samples totally, but five of them are irrelevant and thus only 83 samples were used.
[2] ALL_AML in Dataset Two and Leuk in Dataset One are the same dataset in actual.
[3] MLL dataset was retrieved from http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?paper_id=63.

**Pseudocode 2:** GA-RFE: recursive feature elimination with embedded GA

**Input :** Candidate gene set $G$, Maximal chromosome length $L$
**Output:** The optimal gene feature combinations in GA-RFE
Randomly generate the first GA population $P$ from $G$ with chromosome length equal to $L$;
Set $GC$, the list of optimal gene combinations in GA-RFE, to empty;
**do**
  Execute embedded GA (Pseudocode 3) using the population $P$;
  Add the returned gene combinations by GA to $GC$;
  **if** *the current chromosome length* $> 1$ **then**
    Chromosome length drop: each individual in $P$ randomly discard several genes with the number equal to the RFE step;
**while** *the current chromosome length* $\geqslant 1$;
Sort the optimal gene combinations in $GC$ and only preserve the top ranked ones;
Return the optimal gene combinations in $GC$;

**Pseudocode 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** *the maximal evolution times* $T$ *not reached* **do**
  Perform mutation operator;
  Perform crossover operator;
  Fitness calculation of each GA individual by k-fold CV;
  Truncation selection to form the updated $P$;
Sort the GA individuals in $P$ and select the top ones to form $GC$;
Return the updated population $P$ and the optimal gene combinations in $GC$;

one GA-RFE processes at each layer is usually enough. In the experiments, we limited the size of the final selected gene combination in each dataset to below 10 genes. According to the experiment records, in each of the 19 microarray datasets, the running time of MGRFE is commonly between 500 seconds (8.33 minutes) and 900 seconds (15 minutes). The running time has included the whole filter screen and later wraper search processes. Additionally, it is well worthy to mention that the final chosen gene subset in each dataset might be already found by the first GA-RFE process in the first layer of MGRFE, which just costs 2$\sim$3 minutes. More implementation details and experiment results of MGRFE in the 19 datasets are available at https://github.com/Pengeace/MGRFE-GaRFE.

Because Kar *et al.* also employed an evolutionary-computation method PSO, which is similar to GA, to select minimal informative genes in microarray and provided their program running time records on three datasets SRBCT, ALL_AML, and MLL [2], here, we offer a simple running time comparison between their method and MGRFE. Their PSO-based method cost 2.7956, 2.7906 and 7.1488 hours on SRBCT, ALL_AML and MLL respectively to find their optimal gene subsets. In contrast, MGRFE merely used 10.8230, 9.0108 and 8.8739 minutes respectively in the same three datasets and thus showed much higher converge speed. Moreover, according to Tables 4, 5 and 6 in the main manuscript, the gene subsets selected by MGRFE had smaller sizes but higher classification accuracies compared with Kar *et al.*'s method. We noted that Kar *et al.* didn't employ the filter techniques to cut down the feature search space and their binary-coded PSO don't has any explicit feature decline mechanism like RFE.

## S4. Performance of MGRFE in 10-time 10-fold CV

In the main manuscript, the performance of MGRFE on the two large Datasets in 10-time 10-fold cross validation (CV) are shown in the box-plot form. Here, we supplement the detailed mean accuracies ($Mean\ Acc$s) and standard deviations ($S.D.$s) of MGRFE on all the 19 datasets. In each dataset, 10-fold CV is repeated 10 times based on different random seeds. In each 10-fold CV, the mean accuracy value in 10-fold is calculated and recorded. Then after 10 repetitions of the 10-fold CV, the $Mean\ Acc$ and $S.D.$ of MGRFE in a dataset is calculated from the recorded total 10 mean accuracy values.

## S5. The gene probes selected by MGRFE

The gene probes finally selected by MGRFE on all the 19 datasets are listed in Table 4. These differentially

TABLE 3
$Mean\ Acc$ and $S.D.$ of MGRFE on 19 benchmark datasets in 10-time 10-fold CV

| Dataset | DLBCL | Pros | Colon | Leuk | Mye | ALL1 | ALL2 | ALL3 | ALL4 | CNS |
|---|---|---|---|---|---|---|---|---|---|---|
| $Mean\ Acc$ | 0.987 | 0.979 | 0.971 | 0.982 | 0.933 | 0.998 | 0.880 | 0.920 | 0.963 | 0.980 |
| $S.D.$ | 0.007 | 0.003 | 0.012 | 0.007 | 0.011 | 0.004 | 0.013 | 0.007 | 0.010 | 0.007 |
| Dataset | Lym | Adeno | Gas | Gas1 | Gas2 | T1D | Stroke | SRBCT | MLL | |
| $Mean\ Acc$ | 1.000 | 1.000 | 1.000 | 0.974 | 1.000 | 0.897 | 1.000 | 1.000 | 0.997 | |
| $S.D.$ | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 | 0.014 | 0.000 | 0.000 | 0.006 | |

Note that the $Mean\ Acc$ and $S.D.$ of MGRFE in dataset ALL_AML are same as the records in Leuk for these two are the same dataset in actual.

TABLE 4
The gene probes finally selected by MGRFE on the 19 microarray datasets

| Dataset | Probe number | Gene probes |
|---|---|---|
| DLBCL | 3 | [*X69433_at*, *Z84497_s_at*, *M15205_at*] |
| Pros | 4 | [*37639_at*, *38634_at*, *1909_at*, *37537_at*] |
| Colon | 6 | [*Hsa.36952*, *Hsa.36696*, *Hsa.94*, *Hsa.442*, *Hsa.5226*, *Hsa.5756*] |
| Leuk | 2 | [*M23197_at*, *M31523_at*] |
| Mye | 7 | [*35977_at*, *33130_at*, *31366_at*, *34571_at*, *38013_at*, *1368_at*, *41150_r_at*] |
| ALL1 | 1 | [*38319_at*] |
| ALL2 | 8 | [*37502_at*, *39885_at*, *1291_s_at*, *39408_at*, *1838_g_at*, *819_at*, *31331_at*, *39336_at*] |
| ALL3 | 8 | [*38907_at*, *38478_at*, *34284_at*, *37693_at*, *201_s_at*, *34497_at*, *37809_at*, *41259_at*] |
| ALL4 | 6 | [*39631_at*, *38119_at*, *36795_at*, *36873_at*, *39905_i_at*, *1265_g_at*] |
| CNS | 7 | [*S76475_at*, *M96739_at*, *X64624_s_at*, *X93511_s_at*, *K01911_at*, *S78693_f_at*, *X78565_at*] |
| Lym | 3 | [*GENE3332X*, *GENE3261X*, *GENE1191X*] |
| Adeno | 1 | [*D43636*] |
| Gas | 3 | [*225571_at*, *236118_at*, *237466_s_at*] |
| Gas1 | 3 | [*213125_at*, *41037_at*, *208897_s_at*] |
| Gas2 | 2 | [*212344_at*, *210766_s_at*] |
| T1D | 7 | [*1566232_at*, *215728_s_at*, *215612_at*, *226585_at*, *239474_at*, *219870_at*, *244223_at*] |
| Stroke | 4 | [*1567009_at*, *240084_at*, *239389_at*, *233835_at*] |
| SRBCT | 5 | [*245330.0*, *784257.0*, *43733.0*, *784224.0*, *295985.0*] |
| MLL | 3 | [*38242_at*, *37710_at*, *1389_at*] |

expressed genes could be potential biomarker candidates that are useful to related phenotype researches.

## S6. THE STATISTICALLY SIGNIFICANT GENES IN *t*-TEST

In *t*-test, this study adopt the widely used $p = 0.05$ significance threshold to select the differentially expressed genes with *p*-values lower than the threshold. The chosen of *p*=0.05 has also been experimentally validated by the sample distribution condition of 17 binary classification datasets and our final experiment results on these datasets. Table 5 illustrates the number of significant genes with *p*-value less than 0.05 in the *t*-test. From Table 5 we can note that, *p*=0.05 is a relatively accommodative condition on 17 binary-class datasets, which can not only identify the most differentially expressed genes, but also avoid the inappropriately exclusion of too many genes.

## S7. COMPARE OTHER FILTER METHODS WITH THE *t*-TEST AND MIC COMBINATION

In the feature space reduction stage, the study used *t*-test and MIC for their efficiency and convenience in gene filtering process. The *t*-test has been widely used and validated for detecting differentially expressed genes in microarray [5], [6], [7]. But *t*-test has limitation in dealing with multi-class dataset for multivariate *t*-test can't be performed directly. The recently

TABLE 5
Number of statistically significant features with *t*-test-based *p*-values less than 0.05 on 17 binary classification datasets.

| Dataset | DLBCL | Pros | Colon | Leuk | Mye | ALL1 | ALL2 | ALL3 | ALL4 |
|---|---|---|---|---|---|---|---|---|---|
| Significant features | 2632 | 5061 | 594 | 2449 | 1720 | 4387 | 644 | 571 | 1279 |
| Total features | 7129 | 12625 | 2000 | 7129 | 12625 | 12625 | 12625 | 12625 | 12625 |
| Dataset | CNS | Lym | Adeno | Gas | Gas1 | Gas2 | T1D | Stroke | |
| Significant features | 334 | 804 | 1799 | 8260 | 16454 | 15601 | 10159 | 5569 | |
| Total features | 7129 | 4026 | 7457 | 22645 | 22283 | 22283 | 54675 | 54675 | |

TABLE 6
The performance comparison among different filter method combinations by 5-fold cross validation

| Filter Methods | Dataset | Genes | $Sn$ | $Sp$ | $Acc$ | $Avc$ | $MCC$ | $AUC$ |
|---|---|---|---|---|---|---|---|---|
| *t*-test+MIC | Adeno | 1 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | Gas1 | 3 | 0.986 | 0.973 | 0.980 | 0.980 | 0.961 | 0.99 |
| | Pros | 4 | 0.980 | 0.982 | 0.981 | 0.981 | 0.963 | 0.98 |
| | DLBCL | 3 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | Leuk | 2 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | CNS | 7 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| Anova+FC | Adeno | 2 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | Gas1 | 3 | 0.987 | 0.973 | 0.980 | 0.980 | 0.960 | 0.979 |
| | Pros | 4 | 0.980 | 0.982 | 0.981 | 0.981 | 0.963 | 0.982 |
| | DLBCL | 3 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | Leuk | 4 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | CNS | 8 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| Volcano plot+MIC | Adeno | 1 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | Pros | 4 | 0.980 | 0.982 | 0.980 | 0.981 | 0.963 | 0.968 |
| | DLBCL | 3 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | Leuk | 2 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

proposed MIC shows excellent performance in detecting a wide range of associations in large datasets including microarray [1], [8], and MIC can cope with multi-class dataset. Thus, we combined *t*-test and MIC to complete the feature screen task.

For the execution order, we perform the *t*-test first and then MIC. By the $p$=0.05 significance threshold in the *t*-test, we can quickly find the statistically significant genes, which could notably reduce the gene feature range. Besides, it has been noticed that the MIC calculation is kind of time-consuming compared with *t*-test, thus it is suitable to perform *t*-test first to decrease the gene number.

We also compared the performance of other filter method combinations with the *t*-test+MIC combination, the result are shown in Table 6.

Firstly, the combination of first Anova then Fold change (FC), Anova+FC. The experiment was carried on 3 balanced datasets (Adeno, Gas1 and Pros) and 3 imbalanced datasets (DLBCL, Leuk, and CNS) using 5-flod cross validation. For Anova, the *p*-value threshold was also set as 0.05 as in *t*-test. From Table 6, it can notice that with the combination of Anova+FC, the sizes of finally selected genes are 2, 4, and 8 on datasets Adeno, Leuk and CNS, respectivily. But by the *t*-test+MIC combination, simply 1, 2 and 7 genes are needed to achieve the same performance on the 3 datasets. On the rest of 3 datasets, the two filter method combinations have similar performance. Thus, the filter method combination of Anova+FC is little inferior to the combination of *t*-test+MIC in finding the minimal discriminative gene subset.

Secondly, the combination of first "volcano plot" then MIC, Volcano plot+MIC. The "volcano plot" can combine the advantages of *t*-test and fold-change, thus we use the "volcano plot" to replace the *t*-test. The experiments were performed on 2 balanced datasets (Adeno and Pros) and 2 imbalanced datasets (DLBCL and Leuk) by 5-fold cross validation. According to experiment results in Table 6, these two methods select same size of genes and achieve similar performance on all the tested 4 datasets. In the experiments, we

noted one defect of "volcano plot" for gene selection in a range of different microarray datasets. When we use the "volcano plot" to selected informative genes, in each microarray dataset, we need to hand-tune the $p$-value threshold in $t$-test and fold-change threshold value in FC to obtain the satisfactory result. For example, on dataset Adeno, there are 894 informative genes have FC value larger or equal to 2; but on dataset Pros, the highest FC value in all genes is just 1.46. Thus, for different datasets, the threshold values in "volcano plot" should be different. In fact, for each of the tested 4 datasets, the threshold values in "volcano plot" have been hand-tuned individually and the finally assigned threshold values vary among different datasets. This situation pose difficulty for building automatically application on a wide range of microarray datasets. In contrast, the $t$-test has the consistent $p$-value setting in all the 17 datasets in experiments and shows more convenience.

To conclude, the filter method combination of $t$-test+MIC are more efficient and convenient than the Anova+FC or Volcano plot+MIC combinations for the feature range reduction task in our study.

## REFERENCES


[1] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, and F. Zhou, "Mctwo: a two-step feature selection algorithm based on maximal information coefficient," *BMC bioinformatics*, vol. 17, no. 1, p. 142, 2016.

[2] S. Kar, K. D. Sharma, and M. 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*, vol. 42, no. 1, pp. 612–627, 2015.

[3] E. Jones, T. Oliphant, and P. Peterson, "Scipy: Open source scientific tools for python," 2014.

[4] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, "minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers," *Bioinformatics*, vol. 29, no. 3, pp. 407–408, 2013. [Online]. Available: ⟨GotoISI⟩://WOS:000314892000022

[5] X. Q. Cui and G. A. Churchill, "Statistical tests for differential expression in cdna microarray experiments," *Genome Biology*, vol. 4, no. 4, 2003. [Online]. Available: ⟨GotoISI⟩://WOS:000182696200003

[6] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012. [Online]. Available: ⟨GotoISI⟩://WOS:000304147000018

[7] N. Sato, I. M. Sanjuan, M. Heke, M. Uchida, F. Naef, and A. H. Brivanlou, "Molecular signature of human embryonic stem cells and its comparison with the mouse," *Developmental Biology*, vol. 260, no. 2, pp. 404–413, 2003. [Online]. Available: ⟨GotoISI⟩://WOS:000184946000010

[8] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[9] G. Wang, N. Hu, H. H. Yang, L. Wang, H. Su, C. Wang, R. Clifford, E. M. Dawsey, J.-M. Li, T. Ding *et al.*, "Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china," *PloS one*, vol. 8, no. 5, p. e63826, 2013.