# RIFS: fuse two straws into a diamond

Yuting Ye\*, Ruochi Zhang\*, Weiwei Zheng, Shuai Liu, Fengfeng Zhou#.

College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China.

\* These authors contribute equally to this study.

# Corresponding author: Fengfeng Zhou, Email: FengfengZhou@gmail.com, or ffzhou@jlu.edu.cn. Website: <http://www.healthinformaticslab.org/ffzhou/>.

## Abstract

The advent of big data era has imposed major challenges for the machine learning researchers. Biomedical OMIC research is one of these big data areas and has changed the biomedical research drastically. But the high cost of data production and difficulty in participant recruitment introduce the paradigm of “large p small n” into the biomedical researches. Feature selection is usually employed to reduce the high number of biomedical features, so that a stable data-independent classification or regression model may be achieved. This study randomly changes the starting feature of the widely-used incremental feature selection (IFS) strategy and selects the best feature subset that may be ranked low by the statistical association evaluation algorithms, *e.g.* t-test. The hypothesis is that two low-ranked features may be orchestrated to achieve a good classification performance. The proposed Randomly re-started Incremental Feature Selection (RIFS) algorithm demonstrates both higher classification accuracy and smaller feature number than the existing algorithms. RIFS also outperforms the existing diagnosis model for the prostate malignancy with a better accuracy and a smaller number of features.

**Keywords:** Feature selection, filter algorithm, incremental feature selection (IFS), random re-start, *k*-fold cross validation.

## Introduction

Modern biological technologies are rapidly revolutionized and improved in the recent years, and the biological OMIC data has been accumulated at an accelerated speed [2](#_ENREF_2). Human complex disease like cancers and cardiovascular diseases are known to be associated with more than one genetic factor [3](#_ENREF_3),[4](#_ENREF_4) and the classic single-factor correlation analysis tends to detect low-frequency statistically-significant factors [1](#_ENREF_1). So the existing complex disease diagnosis panels usually use the genetic information of multiple genes [5](#_ENREF_5),[6](#_ENREF_6).

Transcriptome is the orchestrated result of both genetic and environmental factors, and is widely used in the existing disease diagnosis technologies [7](#_ENREF_7). A gene’s transcriptional efficiency is known to be associated with codon usage patterns [8](#_ENREF_8), and its promoter binding affinity with the transcription regulators [9](#_ENREF_9). The environmental factors including diet and air quality may also impact a gene’s activities through the signal transduction machinery as well as the inheritable epigenetic mechanisms [10](#_ENREF_10),[11](#_ENREF_11).

The development of a disease diagnosis panel relies on the efficiency of the feature selection technologies [12](#_ENREF_12). A biological OMIC technology generates thousands or even millions of data entries for a single sample, but a biomedical project seldom recruited more than 1,000 samples due to various limitations, *e.g.* cost and patient availability [13](#_ENREF_13). The biomarker screening procedure may generate the overfit models due to the paradigm of “large *p*, small *n*”, where *p* and *n* are the numbers of features and samples, respectively [14](#_ENREF_14),[15](#_ENREF_15). Besides the aforementioned statistical reason, it is also known that not all the genes are biologically involved in a given disease onset and development processes [16](#_ENREF_16).

It is a computationally infeasible task to find a global optimal feature subset within a reasonable period of time [17](#_ENREF_17) and the existing feature selection algorithms may be roughly grouped as filter and wrapper approximate algorithms [18](#_ENREF_18),[19](#_ENREF_19). A filter algorithm evaluates each feature’s association with the class label using a statistical significance measurement [20](#_ENREF_20). Filter algorithm has been widely used in many biomedical biomarker screening projects due to its linear time requirement, and sometimes is the only choice for large datasets like SNP and methylation polymorphisms [21](#_ENREF_21),[22](#_ENREF_22). But a filter algorithm only ranks the features by their individual associations with the class labels, and the user is responsible for choosing the number of top-ranked features [23](#_ENREF_23). A wrapper algorithm evaluates each heuristically selected feature subset using a classification algorithm, and tends to achieve better classification performance than the filters since a wrapper algorithm directly optimizes the target classification algorithm. A wrapper usually runs much slower than a filter algorithm, due to its consideration of inter-feature relationships [24](#_ENREF_24).

This study proposed a modified incremental feature selection strategy for the filter algorithms. An Incremental Feature Selection (IFS) strategy evaluates the classification performance of the top-*k*-ranked features iteratively for *k*∈(1, 2, …, *n*), where *n* is the total number of features. IFS usually stops at the first observation of performance decrease [18](#_ENREF_18),[25](#_ENREF_25). This study proposes an IFS strategy by selecting features incrementally from a randomly-selected starting feature, and selected the best solution from multiple Randomly re-started IFS (RIFS) procedures. The comparison with the existing filters and wrappers demonstrates that RIFS outperforms them by both higher classification accuracies and smaller feature numbers.

## Material and Methods

### Binary classification problem

This study evaluates a feature subset using the binary classification performance. A binary classification problem has two groups of samples, *i.e.* the Positive (*P*) and Negative (*N*) samples [18](#_ENREF_18),[26](#_ENREF_26). *P* and *N* are also used to denote the numbers of positive and negative samples. The binary classification problem is the simplest classification model, and most of the classification algorithms can solve this problem heuristically. This problem is also the most widely adopted problem setting for biomedical researchers, *e.g.* the problem of disease versus control in the Genome-Wide Association Study (GWAS) [27](#_ENREF_27), and the samples of two phenotypes in the clinical survival analysis [28](#_ENREF_28), *etc*.

### Two groups of feature selection algorithms

This study compares the proposed algorithm with two major groups of feature selection algorithms, *i.e.* filters and wrappers [29-31](#_ENREF_29). Three filters, *i.e.* T-test based ranking (Trank) [32](#_ENREF_32), false positive classification rate (FPR) [33](#_ENREF_33), and Wilcoxon-test based ranking (Wrank) [34](#_ENREF_34), are evaluated when they select the same numbers of features as the proposed algorithm in this study. Wrappers can directly recommend a list of features without the user-determined number of features [35](#_ENREF_35). So this study investigated both the classification performances and the numbers of features for these feature selection algorithms.

### Performance measurements

A binary classification algorithm optimizes the parameters of a model and predicts that a new sample belongs to the positive (*P*) or negative (*N*) group. The sizes of the positive and negative groups are denoted as *P* and *N*, respectively. A positive sample is defined as a true positive or false negative one, if it is predicted as positive or negative. And a negative sample is defined as false positive or true negative, if its prediction is positive or negative. The numbers of true positives, false negatives, false positives and true negatives are denoted as *TP*, *FN*, *FP* and *TN*, respectively. The binary classification performance is evaluated by the following measurements, as similar in [18](#_ENREF_18). Sensitivity (*Sn*) and specificity (*Sp*) are defined as the percentages of correctly predicted positive and negative samples, *i.e.* *Sn*=*TP*/(*TP*+*FN*) and *Sp*=*TN*/(*TN*+*FP*). The overall accuracy (*Acc*) is defined as *Acc*=(*TP*+*TN*)/(*TP*+*FN*+*TN*+*FP*).

Five representative classification algorithms were evaluated on the datasets, and the maximal accuracy achieved by these algorithms for a dataset was defined as *mAcc*, the performance measurement of a given feature subset. Support Vector Machine (SVM) is a widely used binary classification algorithm. K Nearest Neighbors (KNN) algorithm is an intuitive distance based classification algorithm. Decision Tree (DTree) will generate an easy-to-interpret classifier. And Naïve Bayesian classifier (NBayes) assumes that all the features are independent to each other. Logistic Regression (LR) trains a linear classification function, which may suggest the weights of the chosen features.

All the algorithms were tested under two major Python releases, *i.e.* 2.7.13 and 3.6.0, using the default parameters.

### Biomedical datasets

The proposed algorithm in this study was compared with the existing algorithms using 17 datasets, as similar in [19](#_ENREF_19). Each of the 17 datasets has two class labels, *i.e.* a binary classification problem. Six transcriptome datasets, *i.e.* DLBCL [36](#_ENREF_36), Pros [37](#_ENREF_37), ALL [38](#_ENREF_38), CNS [39](#_ENREF_39), Lym [40](#_ENREF_40) and Adeno [41](#_ENREF_41), were publicly available at the Broad Institute Genome Data Analysis Center. The two datasets *Colon* [42](#_ENREF_42) and *Leuk* [43](#_ENREF_43) were provided in the R/Bioconductor packages *colonCA* and *golubEsets*, respectively. The dataset ALL modeled as four binary classification datasets, *i.e.* ALL1/ALL2/ALL3/ALL4, based on different phenotype annotations, as described in Table 1. Five more recent datasets, *i.e.* Mye (accession: GDS531) [44](#_ENREF_44), Gas (accession: GSE37023) [45](#_ENREF_45), Gas1/Gas2 (accession: GSE29272) [46](#_ENREF_46), T1D (accession: GSE35725) [47](#_ENREF_47) and Stroke (accession: GSE22255) [48](#_ENREF_48), were publicly available at the NCBI Gene Expression Omnibus (GEO) database. The raw data from the NCBI GEO database were normalized into the gene expression matrix with the default parameters of the RMA algorithm [49](#_ENREF_49), and all the other datasets were downloaded as the normalized data matrix.

**Table 1. Summary information of the 17 binary classification datasets.** Datasets 1-15 are the cancer transcriptomes, while the last two are transcriptome datasets of type I diabetes and stroke, respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Dataset | Samples | Features | Summary |
| 1 | DLBCL | 77 | 7,129 | DLBCL (58) vs follicular lymphoma (19) |
| 2 | Pros | 102 | 12,625 | prostate cancer (52) vs control (50) |
| 3 | Colon | 62 | 2,000 | colon cancer (40) vs normal (22) |
| 4 | Leuk | 72 | 7,129 | ALL (47) vs AML (25) |
| 5 | Mye | 173 | 12,625 | presence (137) vs absence (36) of focallesions of bone |
| 6 | ALL1 | 128 | 12,625 | B-cell (95) vs T-cell (33) ALL |
| 7 | ALL2 | 100 | 12,625 | ALL with (65) vs without (35) relapse |
| 8 | ALL3 | 125 | 12,625 | ALL with (24) vs without (101) multidrug resistance |
| 9 | ALL4 | 93 | 12,625 | ALL with (26) and without (67) the t(9;22) chromosome translocation |
| 10 | CNS | 60 | 7,129 | medulloblastoma survivors (39) vs treatment failures (21) |
| 11 | Lym | 45 | 4,026 | germinalcentre (22) vs activated B-like DLBCL (23) |
| 12 | Adeno | 36 | 7,457 | colon adenocarcinoma (18) vs normal (18) |
| 13 | Gas | 65 | 22,645 | gastric cancer (29) vs non-malignants (36) |
| 14 | Gas1 | 144 | 22,283 | non-cardia gastric cancer (72) vs normal (72) |
| 15 | Gas2 | 124 | 22,283 | cardia gastric cancer (62) vs normal (62) |
| 16 | T1D | 101 | 54,675 | T1D (57) vs control (44) |
| 17 | Stroke | 40 | 54,675 | ischemic stroke (20) vs control (20) |

A recent study proposed that methylomes outperformed transcriptomic profiles in separating prostate cancers from the controls [50](#_ENREF_50). We demonstrated that a delicate selection of transcriptomic features may also achieve a similarly good classification model compared with methylomes. The dataset GSE55599 was downloaded from the Gene Expression Omnibus (GEO) database [51](#_ENREF_51). The binary classification problem worked on the 32 prostate carcinoma samples and 10 benign prostatic hyperplasia samples. Each sample has 47,231 probesets, *i.e.* features.

### RIFS, a randomly re-started incremental feature selection algorithm

The incremental feature selection algorithm was modified to have a start position *k* and the consecutive performance decreasing cutoff *D*, which is denoted as the algorithm *sIFS*(*k*, *D*). Given a binary classification problem with *n* features and *m* samples, the features are ranked based on their association significance with the binary class labels. A feature’s association significance with the class label is measured by the statistical significance *P value* of the t-test [32](#_ENREF_32). Features are denoted as *fi*, *i*∈{1, 2, …, *n*}, based on their ranks. The algorithm will consecutively add the next feature to the feature subset until the binary classification accuracy decreases for consecutively *D* times.

**Algorithm.** *sIFS*(*k*, *D*)

**Input.** The list of ranked features {*f*1, *f*2, …, *fn*}, and the start position *k*. The consecutive decreasing cutoff is *D*.

**Begin.**

FeatureSubset = [1](#_ENREF_1)

BestFS=FeatureSubset

DecreaseTimes = 0

while DecreaseTimes<=D:

if Accuracy(FeatureSubset)>Accuracy(FeatureSubset∪{NextFeature}):

FeatureSubset= FeatureSubset∪{NextFeature}

DecreaseTimes += 1

else:

BestFS=FeatureSubset

DecreaseTimes = 0

return BestFS

**End.**

The Randomly re-started Incremental Feature Selection (RIFS) algorithm is proposed based on the unit algorithm *sIFS*(*k*, *D*). The basic hypothesis is that a summarization of multiple *sIFS*(*k*, *D*) algorithms may generate a feature subset with better classification accuracy than the classical algorithms sIFS(1, 1).

**Algorithm.** *RIFS*(*K*, *D*)

**Input.** The number of start positions for *sIFS*(*k*, *D*) is *K*, and the consecutive decreasing cutoff is *D*.

Begin.

Solution={}

for i in (1..K):

*k*=random()

Solution=Solution∪{*sIFS*(*k*, *D*)}

BestFS=Solution[0]

for *i* in (1..*K*):

if Accuracy(BestFS)<Accuraccy(Solution[*i*]):

BestFS=Solution[*i*]

return BestFS

End.

### Randomly seeded k-fold cross validation

A *k*-fold cross validation (KFCV) strategy is utilized to calculate the overall classification performance. Given a binary classification dataset with positive and negative samples in the subsets *P* and *N*, respectively, KFCV randomly splits *P* and *N* into *k* equally-sized subsets, respectively. *P*={*P*1, *P*2, …, *Pk*} and *N*={*N*1, *N*2, …, *Nk*}. Iteratively, *Pi*∪*Ni* is selected as the testing dataset, and the other samples are used as the training data for a given classification algorithm. The classification performance measurements *Sn*, *Sp* and *Acc* are calculated based on this round of iteration.

Due to that different data splitting will generate different performances, different random seeds will be employed to produce KFCV calculations, and the classification performance measurements are averaged over all the rounds of KFCV experiments.

### Experimental procedure

RIFS was compared with two major classes of feature selection algorithms, *i.e.* three filters and three wrappers. The performance measurement was calculated using 10-fold cross validation, and the classification accuracy was the maximum value *mAcc* of the five classification algorithms, *i.e.* SVM, KNN, DTree, NBayes and LR. The detailed procedure was illustrated in Figure 1.

All the experiments were carried out in an Inspur Gene Server G100, with 256GB memory, 28 Intel Xeon® CPU cores (2.4GHz), and 30TB RISC1 disk space.



**Figure 1. Experimental setting of this work.** 17 datasets were chosen to compare RIFS with three filters and three wrappers, and the classification performances were calculated using 10-fold cross validations.

## Results and Discussion

### Two optimization rules for the IFS strategy

This study proposes two hypothetical modifications of the Incremental Feature Selection (IFS) strategy [18](#_ENREF_18),[25](#_ENREF_25),[52](#_ENREF_52) based on the experiment data, as shown in Figure 2. The optimization goal is to maximize the binary classification accuracy using the selected feature subset.

1. Randomly re-start the IFS strategy. We generalize the classical version of IFS as the IFS(*i*), which chooses a subset of consecutively ranked features starting from rank *i*. Our hypothesis is that there may exist a feature subset IFS(*i*) with a classification accuracy better than IFS(0). This is supported by the two examples in Figure 2 (a) and (b). The two features with ranks 31 and 32 achieved 79.5% in the overall accuracy for the dataset ALL2, better than 77.0% in accuracy for the top two ranked features by IFS(0), as shown in Figure 2. The four features with ranks 443, 444, 445 and 446 outperformed the top four ranked features by 1.3% in accuracy for the dataset ALL3. Their statistical significances measured in *P-values* are 0.036990, 0.036994, 0.037100 and 0.037105 for these four low-ranked features, respectively. So the IFS strategy may be improved by randomly re-started. (此处用到单遍的多倍交叉验证)



**Figure 2. Two examples of IFS(*i*) perform better than IFS(0).** (a) 2 features starting from the rank *i*= 31 for the dataset ALL2. (b) 4 features starting from the rank *i*=443 for the dataset ALL3.

2. Stop at one decrease does not work well**.** We also observe that a big increase in the classification performance may be achieved with adding two consecutively ranked features, even there is a decrease with adding one of them, as demonstrated by Figure 2 (c) and (d). The feature screening process IFS(37) got the first accuracy decrease (1.5%) for the dataset Colon when adding the fourth feature (ranked 40), but achieved an increase (1.4%) in accuracy even compared with the situation before adding the fourth feature, as shown in Figure 3 (a). Another case was the feature screening process IFS(757) for the dataset T1D, as shown in Figure 3 (b). The integration of the third feature decreased the accuracy by 1.3%, but the next feature (ranked 760) increased the accuracy by 5.4%, which was also higher than the two consecutively ranked features 757 and 758 by 4.1% in accuracy. So the stop strategy of IFS(*i*) needs to tolerate at least one accuracy decrease. (此处用到单遍的多倍交叉验证)



(a) (b)

**Figure 3. Two examples of classification performance fluctuations.** (a) Accuracy curve of IFS(37) for the dataset Colon. (b) Accuracy curve of IFS(757) for the dataset T1D.

### How many starting points are enough for most datasets?

We investigated the best number of starting points for RIFS using the four datasets, ALL1/ALL2/ALL3/ALL4, as shown in Figure 4. RIFS was set to stop if three consecutive tries do not increase the classification performances. Six choices of the numbers of starting points were tested, *i.e.* *pStartingPercentage*=10%, 20%, 30%, 40%, 50% and 60% of the total feature number, respectively. The classification algorithms achieved 100% in *mAcc* for all the six values of the parameter *pStartingPercentage* on the dataset ALL1. It seems that the dataset ALL1 is easy to separate, and some other algorithms also achieved 100% in *mAcc*, as demonstrated in [19](#_ENREF_19). The measurement *mAcc* for the dataset ALL2 was improved from 79.9% to 80.6% when the parameter *pStartingPercentage* was increased from 10% to 60%, and the maximum value 80.6% was achieved after *pStartingPercentage*=40%. The measurement *mAcc* remained 88.3% for all different values of *pStartingPercentage* for the dataset ALL3. And the best *mAcc*=94.9% was achieved after *pStartingPercentage*=50%. So the default value 50% for set for the parameter *pStartingPercentage*. (此处用到单遍的多倍交叉验证)



**Figure 4. How many starting points are enough.** The maximum accuracy is calculated for each of the four datasets, *i.e.* ALL1/ALL2/ALL3/ALL4, with different percentages of all the features as the starting points.

### How much tolerance for consecutive performance decreases is enough?

A greedy feature selection algorithm tends to stop when the optimization goal is decreased during the feature screening process, *e.g.* the classical IFS strategy. Our hypothesis is that after a minor decrease in the classification performance, the next screening step may achieve a much better overall performance improvement, as supported by the two examples in Figure 2 (c) and (d).

We evaluated the RIFS stopping criteria *pStoppingDepth*=1, 2, 3, 4 and 5, *i.e.* RIFS stops when *pStoppingDepth* consecutive performance decreases are detected. The four datasets ALL1/ALL2/ALL3/ALL4 were chosen for the evaluation. The parameter *pStartingPercentage* was set to 10%, 20%, 30%, 40% and 50%, and the performance measurement is *mAcc* by the 10-fold cross validation strategy. Figure 5 demonstrated that RIFS achieved the best *mAcc* when *pStoppingDepth* reached 4 for all the four datasets using the 5 values of *pStartingPercentage*. So the experimental data supported our hypothesis that it may not be the best choice to stop immediately after one performance decrease was detected. And the default value of *pStoppingDepth* was chosen as 4. (此处用到单遍的多倍交叉验证)



**Figure 5. How many steps are tolerated without performance improvements.** The classification performance is measured in *mAcc*.

A comprehensive evaluation of RIFS with the default parameter values *pStartingPercentage*=50% and *pStoppingDepth*=4 was carried out for all the 17 transcriptome datasets, as shown in Figure 6. RIFS achieved at least 0.804 in *mAcc* for these datasets, and even achieved 1.000 in *mAcc* for 6 out of the 17 datasets. The following sections will compare RIFS with the existing feature selection algorithms by the performance measurement *mAcc*. (此处用到20遍的多倍交叉验证)



**Figure 6. The classification performances of RIFS on the 17 transcriptome datasets.** The measurement mAcc is used as the vertical axis, and the horizontal axis lists the 17 datasets. The detailed mAcc values are also given on the top of each column.

### Comparison with filters on the 17 transcriptomes

RIFS with the default parameter values *pStartingPercentage*=50% and *pStoppingDepth*=4 was compared with the filter algorithms. A filter algorithm assumes that features are independent to each other, and evaluates the association of a feature with the class label independently. So users need to determine how many features will be chosen, after all the features are evaluated and ranked by the filer algorithm. In order to conduct a fair comparison, if RIFS selects *k* features, this study selects the same number *k* of top-ranked features evaluated by a filter. 10-fold cross validation strategy was employed to calculate the binary classification performances of RIFS and the three filter algorithms, *i.e.* TRank, FPR, and WRank. 【你没有做多遍KFCV吗？譬如“20遍10-fold cross validations、然后取平均值和方差”？】【我问的不是“做了多倍交叉验证”，而是“是否做了多遍的多倍交叉验证”！以及做了几次“多倍交叉验证”？做完了如何得到现在画图的数值的？】



**Figure 7. Performance comparison of RIFS with three existing filter algorithms.** The vertical axis is the performance measurement *mAcc*, and the horizontal axis gives the dataset names. Since all the filter algorithms select the same number of features as RIFS, only the numbers of features by RIFS are shown.

RIFS performed the best compared with the three filter algorithms on all the 17 datasets, as shown in Figure 7. The four datasets Adeno, ALL1, Lymp and Stroke, seem to be easy to be separated, since three feature selection algorithms including RIFS achieved 100% in *mAcc*. RIFS outperformed the three filter algorithms on all the other 13 datasets. And CNS seems to be a difficult binary classification dataset. RIFS achieved 87.4% in *mAcc*, and improved the other filter algorithms by at least 11.6% in *mAcc*. RIFS usually selects no more than 10 features, and the maximal number of features selected by RIFS was 27 for the dataset Myel.

The experimental data suggests that an orchestration of low-ranked features may achieve very good classification performances even for the difficult datasets like CNS and ALL2. No filter algorithms achieved *mAcc* better than 76.0%, and RIFS only achieved 80.4% and 87.4% in *mAcc* for the datasets ALL2 and CNS, respectively.. This provides an additional piece of evidence for the rule “Randomly re-start the IFS strategy” of RIFS. (此处用到20遍的多倍交叉验证)

### Comparison with wrappers on the 17 transcriptomes



**Figure 8. Performance comparison of RIFS with three existing wrapper algorithms.** The vertical axis is the performance measurement *mAcc*, and the horizontal axis gives the dataset names. The table under the line plots gives the numbers of features selected by the four algorithms.

RIFS with the default parameters *pStartingPercentage*=50% and *pStoppingDepth*=4 performed the best compared with the three wrapper algorithms, *i.e.* Lasso, RF and Ridge, on all the transcriptome datasets except for Myel, as shown in Figure 8. RIFS achieved 100% in *mAcc* for 6 out of the 17 datasets, and its average *mAcc* is 94.5%. The next best algorithm based on the average *mAcc* is Lasso. Lasso achieved the average *mAcc* 90.4% and *mAcc*=100% for the three datasets ALL1, Leuk and Lymp. Lasso was also the only wrapper algorithm outperforming RIFS with 2.4% in *mAcc* on the dataset Myel. Except for this case, RIFS performed better than all the wrapper algorithms on all the datasets, and achieved an average improvement 3.5% in *mAcc* for the best of the three wrapper algorithms on the 17 transcriptomic datasets.

Another performance measurement for a feature selection algorithm is the number of features selected by the algorithm. Besides the excess consumption of computational power in training and predicting by a classification model with a large number of features, the overfitting problem is also inevitable to be fixed [53](#_ENREF_53). Due to the high data production cost in the biomedical area, the number of samples is usually much smaller than that of features in a biomedical dataset [54](#_ENREF_54). And the final clinical deployment of a classification model has a cost positively correlated with the number of features in the model. So a biomedical classification model with a higher accuracy and a smaller number of features is preferred in the clinical settings.

RIFS recommended a smaller list of features for training classification models with higher accuracy except six cases, as shown in the table under the line plot in Figure 8. Both RIFS and Lasso recommended 11 features for the dataset colon, but RIFS outperformed Lasso with an improvement 3.3% in *mAcc*. Lasso selected the same number of features as RIFS for the dataset Lymp, and both achieved 100% in *mAcc*. And Lasso outperformed RIFS with an improvement 2.4% in *mAcc* on the dataset Myel. For the three datasets ALL3, Gas1 and T1D, Lasso recommended fewer features than RIFS, but RIFS achieved improvements in *mAcc* 6.7%, 1.5% and 11.4%, respectively. Overall, RIFS recommended an average number of features 8.882, while the three wrapper algorithms chose 25.706, 54.706, and 8428.353 features, respectively. (此处用到20遍的多倍交叉验证)

Transcriptome performs similarly well with methylome

RIFS was employed to screen the transcriptomes of prostate samples, and detected two features with 100% discrimination accuracy of prostate cancers. A recent study suggested that top-ranked methylome features outperformed the top-ranked transcriptome features by at least 5.7% in the measurement Area Under the ROC Curve (AUC) [50](#_ENREF_50). A combination of three methylome features achieved 100% in both the overall accuracy and AUC, while three expression features only achieved 0.978 in AUC. RIFS detected two features ILMN\_1708743 and ILMN\_1727184 as the biomarkers to discriminate the samples of prostate carcinoma and benign prostatic hyperplasia, as illustrated in Figure 9. We may see that these two biomarkers can easily separate the 32 prostate carcinoma and 9 benign prostatic hyperplasia samples. There is only one sample which is very close to the prostate carcinoma samples in the two-dimensional plain. And a non-linear-kernel SVM can achieve 100% in *mAcc* for this binary classification problem. (此处用到20遍的多倍交叉验证)



**Figure 9. Dot plot of the two features detected by RIFS.** There is only one benign prostatic hyperplasia sample which is very close to the prostate carcinoma ones.

## Conclusions

RIFS demonstrated a new perspective of feature selection that two individually low-ranked features may work together to make a highly accurate classification model. RIFS can detect features with very good classification performances, by significantly expanding the searching space. RIFS also tries to avoid the local optimal solutions by tolerating more than one classification performance decreases. There is a balance between the running time and the classification performance, but the user has the flexibility of choosing better classification accuracy for a long running time, or an acceptable accuracy within a short period of time.

## Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400) and the startup grant from the Jilin University.

## Author Contributions

FZ conceived the project, designed the experiment and wrote the manuscript. YY coded the initial algorithm in R and completed the performance comparison with the existing algorithms. RZ re-implemented the RIFS algorithm with improved rules and conducted all the performance comparison procedures using Python. A demonstration of how RIFS performs on the methylation data was also carried out by RZ. WZ and SL helped draft and proof-read the manuscript.

## References

1 Atanasovska, B., Kumar, V., Fu, J., Wijmenga, C. & Hofker, M. H. GWAS as a Driver of Gene Discovery in Cardiometabolic Diseases. *Trends Endocrinol Metab* **26**, 722-732, doi:10.1016/j.tem.2015.10.004 (2015).

2 Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol* **13**, e1002195, doi:10.1371/journal.pbio.1002195 (2015).

3 Dai, X., Xiang, L., Li, T. & Bai, Z. Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. *J Cancer* **7**, 1281-1294, doi:10.7150/jca.13141 (2016).

4 Selvaraju, V. *et al.* Diabetes, oxidative stress, molecular mechanism, and cardiovascular disease--an overview. *Toxicol Mech Methods* **22**, 330-335, doi:10.3109/15376516.2012.666648 (2012).

5 Figueroa, J. D. *et al.* Genome-wide interaction study of smoking and bladder cancer risk. *Carcinogenesis* **35**, 1737-1744, doi:10.1093/carcin/bgu064 (2014).

6 Cuperlovic-Culf, M., Belacel, N., Davey, M. & Ouellette, R. J. Multi-gene biomarker panel for reference free prostate cancer diagnosis: determination and independent validation. *Biomarkers* **15**, 693-706, doi:10.3109/1354750X.2010.511268 (2010).

7 Sparano, J. A., Goldstein, L. J., Davidson, N. E., Sledge, G. W., Jr. & Gray, R. TOP2A RNA expression and recurrence in estrogen receptor-positive breast cancer. *Breast Cancer Res Treat* **134**, 751-757, doi:10.1007/s10549-012-2112-7 (2012).

8 Hajjari, M., Khoshnevisan, A. & Behmanesh, M. Compositional features are potentially involved in the regulation of gene expression of tumor suppressor genes in human tissues. *Gene* **553**, 126-129, doi:10.1016/j.gene.2014.10.011 (2014).

9 Jelkmann, W. Control of erythropoietin gene expression and its use in medicine. *Methods Enzymol* **435**, 179-197, doi:10.1016/S0076-6879(07)35010-6 (2007).

10 Yang, C. S. & Wang, H. Cancer Preventive Activities of Tea Catechins. *Molecules* **21**, doi:10.3390/molecules21121679 (2016).

11 Holloway, J. W., Savarimuthu Francis, S., Fong, K. M. & Yang, I. A. Genomics and the respiratory effects of air pollution exposure. *Respirology* **17**, 590-600, doi:10.1111/j.1440-1843.2012.02164.x (2012).

12 Baek, S., Tsai, C. A. & Chen, J. J. Development of biomarker classifiers from high-dimensional data. *Brief Bioinform* **10**, 537-546, doi:10.1093/bib/bbp016 (2009).

13 Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**, A68-77, doi:10.5114/wo.2014.47136 (2015).

14 Sanchez, B. N., Wu, M., Song, P. X. & Wang, W. Study design in high-dimensional classification analysis. *Biostatistics* **17**, 722-736, doi:10.1093/biostatistics/kxw018 (2016).

15 Shujie, M. A., Carroll, R. J., Liang, H. & Xu, S. Estimation and Inference in Generalized Additive Coefficient Models for Nonlinear Interactions with High-Dimensional Covariates. *Ann Stat* **43**, 2102-2131, doi:10.1214/15-AOS1344 (2015).

16 Li, Y. & Patra, J. C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* **26**, 1219-1224, doi:10.1093/bioinformatics/btq108 (2010).

17 Yusta, S. C. Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters* **30**, 525-534 (2009).

18 Guo, P. *et al.* Gene expression profile based classification models of psoriasis. *Genomics* **103**, 48-55, doi:10.1016/j.ygeno.2013.11.001 (2014).

19 Ge, R. *et al.* McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinformatics* **17**, 142, doi:10.1186/s12859-016-0990-0 (2016).

20 Radovic, M., Ghalwash, M., Filipovic, N. & Obradovic, Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **18**, 9, doi:10.1186/s12859-016-1423-9 (2017).

21 Ciuculete, D. M. *et al.* A methylome-wide mQTL analysis reveals associations of methylation sites with GAD1 and HDAC3 SNPs and a general psychiatric risk score. *Transl Psychiatry* **7**, e1002, doi:10.1038/tp.2016.275 (2017).

22 Lin, H. *et al.* Methylome-wide Association Study of Atrial Fibrillation in Framingham Heart Study. *Sci Rep* **7**, 40377, doi:10.1038/srep40377 (2017).

23 Gardeux, V. *et al.* Computing molecular signatures as optima of a bi-objective function: method and application to prediction in oncogenomics. *Cancer Inform* **14**, 33-45, doi:10.4137/CIN.S21111 (2015).

24 Yu, L. & Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research* **5**, 1205-1224 (2004).

25 Chen, W., Ding, H., Feng, P., Lin, H. & Chou, K. C. iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* **7**, 16895-16909, doi:10.18632/oncotarget.7815 (2016).

26 Zhou, F. & Xu, Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**, 2051-2052, doi:10.1093/bioinformatics/btq299 (2010).

27 Chapuis, J. *et al.* Genome-wide, high-content siRNA screening identifies the Alzheimer's genetic risk factor FERMT2 as a major modulator of APP metabolism. *Acta Neuropathol*, doi:10.1007/s00401-016-1652-z (2016).

28 Shirahata, M. *et al.* Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. *Clin Cancer Res* **13**, 7341-7356, doi:10.1158/1078-0432.CCR-06-2789 (2007).

29 Dash, M. & Liu, H. Feature selection for classification. *Intelligent data analysis* **1**, 131-156 (1997).

30 Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research* **3**, 1157-1182 (2003).

31 Liu, H. & Yu, L. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on* **17**, 491-502 (2005).

32 Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519 (2001).

33 Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157-1182 (2003).

34 Liu, W. M. *et al.* Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**, 1593-1599 (2002).

35 Kohavi, R. & John, G. H. Wrappers for feature subset selection. *Artificial intelligence* **97**, 273-324 (1997).

36 Shipp, M. A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8**, 68-74, doi:10.1038/nm0102-68 (2002).

37 Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209 (2002).

38 Chiaretti, S. *et al.* Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771-2778, doi:10.1182/blood-2003-09-3243 (2004).

39 Pomeroy, S. L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436-442, doi:10.1038/415436a (2002).

40 Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511, doi:Doi 10.1038/35000501 (2000).

41 Notterman, D. A., Alon, U., Sierk, A. J. & Levine, A. J. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research* **61**, 3124-3130 (2001).

42 Alon, U. *et al.* 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* **96**, 6745-6750 (1999).

43 Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

44 Tian, E. *et al.* The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *N Engl J Med* **349**, 2483-2494, doi:10.1056/NEJMoa030847 (2003).

45 Wu, Y. H. *et al.* Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut* **62**, 1100-1111 (2013).

46 Wang, G. S. *et al.* Comparison of Global Gene Expression of Gastric Cardia and Noncardia Cancers from a High-Risk Population in China. *Plos One* **8** (2013).

47 Levy, H. *et al.* Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. *Genes and Immunity* **13**, 593-604 (2012).

48 Krug, T. *et al.* TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. *Journal of Cerebral Blood Flow and Metabolism* **32**, 1061-1072 (2012).

49 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264, doi:10.1093/biostatistics/4.2.249 (2003).

50 Paziewska, A. *et al.* DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. *Br J Cancer* **111**, 781-789, doi:10.1038/bjc.2014.337 (2014).

51 Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods in molecular biology* **1418**, 93-110, doi:10.1007/978-1-4939-3578-9\_5 (2016).

52 Chen, L., Zhang, Y. H., Huang, T. & Cai, Y. D. Gene expression profiling gut microbiota in different races of humans. *Sci Rep* **6**, 23075, doi:10.1038/srep23075 (2016).

53 Lumbreras, B. *et al.* Sources of error and its control in studies on the diagnostic accuracy of “‐omics” technologies. *PROTEOMICS-Clinical Applications* **3**, 173-184 (2009).

54 Kosorok, M. R. & Ma, S. Marginal asymptotics for the “large p, small n” paradigm: with applications to microarray data. *The Annals of Statistics* **35**, 1456-1486 (2007).