# Randomly re-started Incremental Feature Selection for the Biomedical Data

Yuting Ye1, [根据实际贡献决定其他作者], Fengfeng Zhou1,2,\*.

1 College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China.

2 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China.

\* 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 had led to various researches about mathematical modeling and analysis. The big data based health diagnosis of cancer markers has broad application prospects and important implications for human health，is becoming an import research focus. Selecting the optimal subset of features from a high -dimensional set using traditional feature selection method is computationally infeasible for the computing time of exhaustive search for the best feature subset grows exponentially with the ultra-dimensions. Feature selection has proven to be an NP-hard problem. In order to analyze huge amounts of big data, a variety of feature selection algorithm emerged. Feature selection is a dimensionality reduction process aimed at weeding out the irrelevant redundant data from a complicated and redundant high dimensional dataset. This work describes a feature selection algorithm RIFS based on IFS(Incremental Feature Selection). It presents the idea of random Re-start. In this work, m randomly generated numbers will be used as the starting position of the ascending sorted feature set. From the starting position, followed by the selection of k feature, as a feature subset, m subsets will be picked out respectively. 3-fold cross validation and 4 classifiers will be used to judge these subsets. In this work, RIFS is applied into 16 datasets. According to the test results, this algorithm performs as well as or better than existing algorithms, the features selected by RIFS appear to have high biomedical correlation with cancer markers, besides, the numbers of selected features are significantly reduced and relatively smaller than these selected by other algorithms. Shown as the numbers of selected features and the performance, RIFS algorithm can correctly select features, effectively remove redundant features.

Keywords: Feature selection ，Random Re-start ，K-fold cross validation，

Wrapper algorithm

## Introduction

At present, due to environmental pollution and other factors, more and more diseases are getting closer to people's lives. Cancer is one of the most terrible diseases that threaten human health. How to prevent and cure cancer has become a hot topic. The discovery and prediction of cancer genes for the prevention and treatment of cancer is of great significance. So how to select cancer related genes is particularly important while domestic and foreign scholars have done a lot of research so far. The following are a number of several representative algorithms, Heuristic Breath first Search (HBSA) Algorithm[Wang Shulin. Wang Ji. Heuristic Breath-First Search Algorithm for Informative Gene Selection Based on Gene Expression Profiles[J]. Chinese Journal of Computers, 2008,31(4):636-649], ARWAH Algorithm [ZHANG Songyao; ZHANG Shaowu. Predicting Lung Cancer Risk Disease Genes Based on Asynchronously Random Walk with Restart in Heterogeneous Networks [J].[Acta Biophysica Sinica](http://sjk11.e-library.com.cn/kns55/loginid.aspx?uid=&p=Navi%2FBridge.aspx%3FLinkType%3DBaseLink%26DBCode%3Dcjfq%26TableName%3DCJFQbaseinfo%26Field%3DBaseID%26Value%3DSWWL), 31(1):33-34],IFS Algorithm [[H Liu](http://xueshu.baidu.com/s?wd=author%3A%28Huan%20Liu%29%20Department%20of%20Information%20Systems%20and%20Computer%20Science%2C%20National%20University%20of%20Singapore%2C%20Kent%20Ridge%2C%20Singapore%2C%20119260&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson).[R Setiono](http://xueshu.baidu.com/s?wd=author%3A%28Rudy%20Setiono%29%20Department%20of%20Information%20Systems%20and%20Computer%20Science%2C%20National%20University%20of%20Singapore%2C%20Kent%20Ridge%2C%20Singapore%2C%20119260&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson) [Incremental Feature Selection](http://xueshu.baidu.com/s?wd=paperuri%3A%2850bbee30dc60e1eecac7044953da1883%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Flink.springer.com%2F10.1023%2FA%3A1008363719778&ie=utf-8)[J].[Applied Intelligence](http://xueshu.baidu.com/s?wd=journaluri%3A%283466212f89f12c3a%29%20%E3%80%8AApplied%20Intelligence%E3%80%8B&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dpublish&sort=sc_cited), 1998, volume 9(3):217-230(14)] , etc. 【参考文献最好找国际期刊上的英文论文，而不是Chinese Journal of Computers等】

Feature selection is a process of selecting the most effective features from the group features to reduce the dimension of the feature space and eliminating redundant or irrelevant sub features from the original feature set. Based on the evaluation criterion of candidate feature subset and subsequent fusion of the learning and classification algorithms, Various published feature selection algorithms can be roughly classified into three types: Filter, Wrapper and Embedded.

## Related work

Filter feature selection has nothing to do with a particular prediction model. This selection strategy assigns a score for each feature based on statistical methods. Here, Filter algorithm measures the association of each feature with the corresponding sample labels. And all features will be sorted according to the measurement. The top-ranked features and feature subset are generally considered to get better accuracy in the following classification algorithms. Filter algorithm does not depend on the specific prediction model, such as the classifier, and the algorithm is of low complexity with good generality,. Therefore, when calculating large-scale data sets, the algorithm can quickly remove a large amount of irrelevant features, and is very suitable to be used as a pre-filter.

[To be summarized：简介三类特征选择算法Filter, Wrapper and Embedded，以及每一个类型的代表性算法]

## Material and Methods

### Binary classification problem and algorithms

Classification is an important issue of data minning, aims at learning a classification function or a classification model which defines the mapping between the dataset and a given set of categories. Classification , in simple terms , divides the data into an existing category based on its characteristics or attributes. The binary classification can be described as follows : input data , namely the Training Set, herein is the feature vector composed of a plurality of sub- features .Each feature of the Training set has a particular corresponding Class Label. Analyzes the input data using the classification model and predict the corresponding classification based on the demonstrated characteristics of the Training set. Four classification algorithms are applied into this paper including Decision Tree(D-Tree) algorithm, Naïve Bayes (N-Bayes)algorithm, Support Vector Machine (SVM) algorithm, K-Nearest Neighbors( KNN) algorithm.

**Decision Tree(D-Tree)**algorithm is an important technology in data mining .By definition , D-Tree is a prediction model relying on the tree structure where the tree structure relying on policy choices. Decision tree is generally divided into three sections , decision nodes, branches and leaf nodes. Headed to the root node, each decision node represents a decision problem or condition and make decisions about which route to take next. Each branch on behalf of the attribute value its parent node proposed. Each leaf node represents the classification result obtained by the different decision nodes which start from the root node along. Throughout the tree, a leaf node represents a classification result , non-leaf nodes represent a problem or condition .

**Bayes classification** is a general term for a class of classification algorithms, which based on the Bayes’ theorem. Naive Bayes (N-Bayes) is the easiest classifier. As the fundamental theorem of Bayes classifier, the Bayes’ theorem solved the common problem of life: known the probability of the event A under the conditions of occurrence of event B, how to obtain the probability of the event exchanged. Here gives the Bayes’ theorem without proof.

【公式的格式：其他公式可以拷贝这个公式过去后修改，使用Microsoft Equation 3.0。

】

The following is the ideological foundation of Naïve Bayes:for the given collection of dataset to be classified and the categories, calculate the probability of each category under the condition and the class lable with the maximum probability will be picked out as the final classification.

(1).Let as a dataset to be classified where each a is one feature of X.

(2).Here is an existing set of categories , calculate

(3). Find the most probable category according to ,the dataset X will be classified to calss which has the maximum probability.

Naive Bayesian classification method is simple, fast , high accuracy , can be applied to large databases .

**Support Vector Machine (SVM)** divide the data points through the construction of one or more high-dimensional hyperplane as a classification boundary.The quality of the boundary is determined by the distance between the nearest data point and the boundary itself. The farther the distance , the lower generalization error .

**K-Nearest Neighbors( KNN)** algorithm is a classic classification algorithm.The easiest way to classify an unknown dataset is to traverse all the data and find a dataset has exactly the same attributes. This approach is not feasible in most cases cause it’s time-consuming and may not find a matched dataset.The main idea of KNN algorithm is to find k nearest records ,and determine the new data according to their main category. KNN classification algorithm can generally be divided into three steps . First calculate the the distance between data set and each training set of data. Then find *k* nearest training data as neighbors. Finally,define the unknown data object class according to the *k* neighbors’ attribute category.

### Performance measurements

Classification algorithms need to configure two categories for an instance in a binary classification in this paper . Common examples of binary classification include predicting whether a patient is suffering from a disease , whether a message is spam.In this work ,two sets of samples are needed to explore a binary classification problem.The number of features in the current sample is n and m respectively. The sample set P and the sample set N, where P is the positive class set (Set Positive), and N is the negative class set (Set Negative). There is now a feature set X , and each feature of X is included in the set P or N , a binary classification is to determine the feature set X belonging to the positive class or negative class according to certain classification algorithms.

There will be four cases for a binary classification problem.If a positive set P is predicted to be positive or negative,then it is marked with TP(ture positive) or FN(false negative) respectively. Accordingly, if a negative set is predicted to be a negative class, called the TN (true negative),or FP(false positive) on the contrary. Sensitivity (Sn), specificity (Sp) and accuracy (Acc) were widely used to measure how well a binary classification model performs [15-17]. The mathematical definition of each index is as follows,Sn=TP/(TP+FN)，Sp=TN/(TN+FP)，Acc=(TP+TN)/(TP+FN+TN+FP). In this issue , subset of features with higher Acc and a smaller number of features will choosed as the optimal feature subset.

### Biomedical datasets

This paper applied 17 binary classification datasets as the basis of classification evaluation. The datasets *Colon*[6] and *Leukaemia* [21] were extracted from the R package *ColonCA* and Bioconductorsix package*golubEsets*, respectively. Another six commonly used datasets *DLBCL*[7], *Prostate*[8], *ALL*[9], *CNS*[10], *Lymphoma*[11]and *Adenoma*[12] were download from the Broad Institute Genome Data Analysis Center, web links http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. Wherein the dataset ALL is further devided into four datasets ALL1,ALL2,ALL3,ALL4. And six datasets *Myeloma*  (accession: GDS531)[13], *Gastric*(accession: GSE37023) [14]. Gastric1/Gastric2 (accession: GSE29272) [15] and T1D(accession: GSE35725)[16] and Stroke (accession: GSE22255) [17],were downloaded from the NCBI Gene Expression Omnibus (GEO) database. Table 1 below shows the number of samples and the number of features in each dataset

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Dataset | Samples | Features |
| 1 | DLBCL | 77 | 7129 |
| 2 | Pros (Prostate) | 102 | 12625 |
| 3 | Colon | 62 | 2000 |
| 4 | Leuk(Leukaemia) | 72 | 7129 |
| 5 | Mye (Myeloma) | 173 | 12625 |
| 6 | ALL1 | 128 | 12625 |
| 7 | ALL2 | 100 | 12625 |
| 8 | ALL3 | 125 | 12625 |
| 9 | ALL4 | 93 | 12625 |
| 10 | CNS | 60 | 7129 |
| 11 | Lym (Lymphoma) | 45 | 4026 |
| 12 | Adeno (Adenoma) | 36 | 7457 |
| 13 | Gas (Gastric) | 65 | 22645 |
| 14 | Gas1 (Gastric1) | 144 | 22283 |
| 15 | Gas2 (Gastric2) | 124 | 22283 |
| 16 | T1D | 101 | 54675 |
| 17 | Stroke | 40 | 54675 |

Table(1) Summary of the 17 binary classification datasets

### Cross Validation

Cross validation technique is conducted to evaluate whether the results of a statistical analysis can be extended to an independent dataset, commonly used to assess the accuracy of a predictive model in practical application . A cross-validation sample dataset need to be divided into two complementary subsets , one subset for training , referred to as the training set; while another subset is called testing set, which is used to verify the validity of the analysis .The research expectation of cross validation is to obtain higher prediction accuracy and lower prediction error. In order to improve the accuracy of cross-validation results ,we generally divide the original sample dataset into multiple different complementary subsets , and then do multiple cross-validation according to different testing sets.

K -fold cross validation , by definition , will split the initial sample dataset into K subsets , taking one subsample as a testing set , and the remaining K-1 subsample is used as a training set .Cross validation is repeated K times, each transform a different subsample as a testing set, to ensure that each sample is verified once. Finally, the average value of the K cross validation results is taken as the verification result. 10- fold cross-validation , for example, the existing sample dataset A will be cut into 10 sub- datasets obtained .Firstly, take the subset as the testing set and the rest 9 subsets as the training set, to get a cross validation result .Then taking as a testing set, the remaining as a training set, get the results of the validation .Similarly, take the next as a testing set, and the rest of the remaining part respectively, as the corresponding training set, as appropriate verification results obtained . After averagingis the final validation result.

In this paper, 3-fold cross validation is applied to evaluate the performance measurements of the classification algorithms mentioned above.

### RIFS algorithm

About the 17 existing high latitude sample datasets of cancer markers, how to extract the cancer related optimal feature subset from a respective original dataset is the issue needs to be solved and Incremental Feature Selection with Random Re-start(RIFS)algorithm is the proposed algorithm to solve this problem .

For each dataset，we can easily calculate the correlation between each feature of cancer and sub-features with higher correlation are more likely to be singled out. But must the optimal feature subset be a combination of sub- features with higher correlation ?Do these sub-ranked features have any contributions to the extraction of optimal feature subset? We therefore make the following conjecture, sub-ranked features may also have good performance. Here put forward the idea of ​​random restart , *M* randomly generated data will be offered as the different starting positions to find the optimal subset .

T-test can infer the probability of difference so it can be used to determine if two sets of data are [significantly](https://en.wikipedia.org/wiki/Statistical_significance) different from each other. P-value as the result of t-test, ranged [0,1], is used to represent the dependence between two variables. Sort the dataset based on the P-value. M randomly generated numbers are respectively defined as the starting positions in the sorted dataset. Beginning with each starting point，find k features in order and form M feature subsets respectively. Four classification algorithms including D-tree, SVM, Nbayes, KNN，are applied to classify these M feature subsets based on the existing Class Label. Using 3-fold cross validation to estimate the accuracy of the classifier and the feature subset with the highest ultimate Acc will be selected as the optimal subset.

Algorithm *.RIFS*

**Input:**(*F*,*C*,*maxnum*) where ,;*maxnum* is the number of the features a biggest featureset contains.

Begin:

Create an array *PvalueFC*[1...*k*] //store *P-value* between all feature and class

Create an array *Subset*[1*..k*] //store subset by the feature ID

numSubset = 1

**for**

PvalueFC [ i ] = t-test (( i ),)

Subset[numSubset]=i

numSubset = numSubset+1

end for

rank the items in Subset[1…numSubset] in the descending order by *PvalueFC*[*Subset*[*i*]]

Create an array *topX*(1,2,…,*x*)

Geneate m random data rd()

Bind topX with rd get Startpositions(1,2…x,)

mAcc = 0;Startp = 0;Countfeature = 0;

**for** i in Startpositions

**for** *h*=1 to *maxnum*

Create an array Subfea**<>**

*tempmAcc* = GetmAcc(*Subfea*);//caculate the Acc of this subfeature

if *tempmAcc* > *mAcc*

mAcc = tempmAcc

*Startp* = *i*

Countfeature = h

**end if**

end for

end for

**return** mAcc;Startp;Countfeature

End

The RIFS algorithm employs function *t-test t*o get *P-value* the between all features and class then ranks these features in the descending order by *P-value*.The RIFS algorithm generates a random dataset *rd* .The function *GetmAcc* is employed to calculate the *Acc* of these feature subsets with numbers in *rd* as start positions in which classifiers and 3-fold cross validation are applied. The variable *tempmAcc* is used to store the result of function *GetmAcc*.IF the value of *tempmAcc* is greater than the current variable *mAcc*，then *mAcc* is replaced by *tempmAcc*, and *Startp* and *Countfeature* as well.

Figure 1 below is the general algorithm process.

Input data数生

DLBCL

CNS

Pros

Colon

Mye

ALL1/2/3/4

Lym

Adeno

Gas/1/2

DLBCL

T1D

Stroke

Other algorithm

RIFS algorithm

Random data

t-test rank

Gene selection

Evaluation

3-fold

cross validation

SVM

KNN

NBayes

D-Tree

Output

Acc

Avc

Sn

Sp

Figure 1 【Inkscape画图，给我原始图SVG格式】

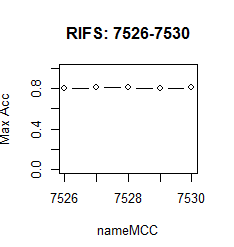
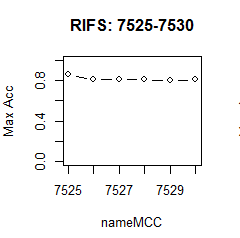
## Results and Discussion

### Three optimization rules

Constantly optimization is needed in order to improve the accuracy and feasibility of the algorithm. Here, three ways of optimization are proposed according to the design ideas of RIFS algorithm above. 【每一个rule给出2个不同数据集上的有较大改进的各1个例子。不用是该数据集上最优的那个特征组合，只要求能够证明：当前优化规则有较好效果。请同时发给我画出这几个例子的数据、代码。】

Save the top-ranking sub features. If directly take the randomly generated data as the starting position of the feature extraction after the ranking of p-value between each features and class, it is easy to ignore the top-ranking features cause the randomly generated data may not contain the top position. Since the features with higher correlation are more likely to be singled out, the ignorance of top-ranking features may cause a great loss of accuracy. Therefore, in the process of the algorithm, we would bind the randomly generated dataset with the top 20 features as shown in the algorithm framework line 12 above.

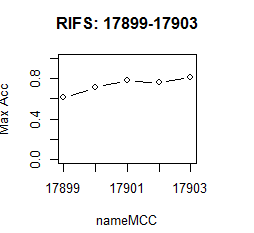
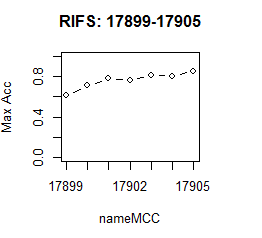
Forward search, expand the search range**.** Take the following RIFS curve in figure 2 (a)as an example. The accuracy of the combination of features ranked 7526th to 7527th in dataset *ALL3*[]showed an increasing state and the highest accuracy is 0.808. So we assumed that adding one or several sub features forwards may come out with a higher accuracy. As shown in figure 2(b)，the feature ranked 7525th has a better performance with the accuracy 0.856. This instance proved that expanding the search forwards may lead to a better performance.

(a) (b)

Figure 2

Expand the search backwards. Take the following RIFS curve in figure 3 (a)as an example. The accuracy of the combination of features ranked 17899th to 17903th in dataset *Gastric2* []showed an increasing state and the highest accuracy is 0.815. The highest point of the curve appears at the end so we assumed that adding one or several sub features backwards may come out with a higher accuracy. As shown in figure 3(b)，the combination of features ranked 17899th to 17899th has a better performance with the value 0.855. This instance proved that expanding the search forwards could lead to a better performance.

(b)

Figure 3

### Evaluation of feature ranking strategies

【测试不同filter算法对RIFS的影响，根据结果推荐一个最佳的filter算法。】【测试数据集就用2个其他算法结果最差的数据集】

Algorithm RIFS is employed to calculate the 17 datasets including DLBCL, CNS, ALL, ect. After optimization, the calculated highest accuracy (mAcc) of dataset DLBCL is 0.987 while the number of features is 9. 5 features are selected from Dataset Prostate with the calculated mAcc-0.951. The mAcc value of dataset Colon is 0.935, and the number of features is 6. The following figure 4 shows the calculated mAcc of the extracted optimal feature subset for the whole 17 datasets

【比较：

1. 基于T-test的RIFS和IFS（请注意每次运行RIFS，需要重新设置随机种子，譬如为：set.seed(0)；
2. 基于Wilcoxon test
3. 基于其他filter算法的比较

请给我提供：画出下面这样柱状图的原始数据、以及画图的代码或EXCEL文件。】

Figure 4

### Evaluation of the initial window size

【测试从随机开始点出发的特征子集窗口大小对RIFS的影响，根据结果推荐一个最佳参数。】【测试数据集就用2个其他算法结果最差的数据集】

### Evaluation of the forward screening depth

【测试从随机特征子集窗口前向扩展对RIFS的影响，根据结果推荐一个最佳参数。就是通常数据集上扩展多远就可以最优了。譬如你可以设置前向搜索100个，取实验数据集上最优结果的前向深度最大值。】【测试数据集就用2个其他算法结果最差的数据集】

### Evaluation of the backward screening depth

【测试从随机特征子集窗口后向扩展对RIFS的影响，根据结果推荐一个最佳参数。就是通常数据集上扩展多远就可以最优了。譬如你可以设置后向搜索100个，取实验数据集上最优结果的前向深度最大值。】【测试数据集就用2个其他算法结果最差的数据集】

**【接下来的性能比较，就都用RIFS的上述推荐默认参数了。】**

### Comparison with filters

【类似于McTwo论文，使用准确性、准确性方差等指标】

### Comparison with wrappers

【类似于McTwo论文，使用准确性、准确性方差等指标】

To illustrate the performance of the algorithm RIFS , this paper also introduced four other existing feature selection algorithms , namely PAM [18], RRF [19], CFS [20], McTwo [21]. To clearly show the performance comparison between the algorithms, the optimum performance data of each algorithm in different datasets will given in the following tables. Table 1 clarifies the mAcc of various algorithms in different datasets while table 2 shows the number of features in these extracted optimal feature sets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | DLB | Pros | Colon | Mye | ALL1 | ALL2 | ALL3 | ALL4 | Lueka |
| RIFS | 0.987 | 0.951 | 0.935 | 0.832 | 1 | 0.77 | 0.856 | 0.871 | 0.972 |
| CFS | 0.987 | 0.949 | 0.919 | 0.899 | 1.000 | 0.837 | 0.838 | 0.948 | 1 |
| PAM | 0.966 | 0.920 | 0.916 | 0.853 | 1.000 | 0.651 | 0.805 | 0.915 | 0.986 |
| RRF | 0.943 | 0.864 | 0.886 | 0.794 | 0.975 | 0.666 | 0.808 | 0.885 | 0.932 |
| McTwo | 0.986 | 0.933 | 0.892 | 0.852 | 1 | 0.716 | 0.822 | 0,894 | 0.993 |
|  | Lym | Aden | Gas | Gas1 | Gas2 | T1D | Strok | CNS |
| RIFS | 0.933 | 1 | 0.969 | 0.972 | 1 | 0.782 | 0.975 | 0.883 |
| CFS | 1.000 | 1.000 | 0.975 | 0.972 | 0.988 | 0.905 | 1.000 | 0.843 |
| PAM | 0.995 | 0.999 | 0.941 | 0.949 | 0.977 | 0.716 | 0.793 | 0.771 |
| RRF | 0.943 | 0.982 | 0.892 | 0.957 | 0.988 | 0.698 | 0.958 | 0.643 |
| McTwo | 0.993 | 1 | 0.977 | 0.953 | 0.992 | 0.795 | 0.843 | 0.828 |

Table 1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | DLBC | Pros | Colo | Mye | ALL1 | ALL2 | ALL3 | ALL4 | Leuk |
| RIFS | 9 | 5 | 6 | 1 | 8 | 3 | 1 | 6 | 8 |
| CFS | 141 | 80 | 53 | 107 | 103 | 56 | 43 | 105 | 39 |
| PAM | 52 | 2 | 14 | 34 | 1 | 1 | 2 | 30 | 20 |
| RRF | 12 | 14 | 16 | 39 | 10 | 31 | 34 | 16 | 29 |
| McTw | 4 | 3 | 6 | 7 | 1 | 2 | 5 | 2 | 2 |
|  | Lym | Aden | Gas | Gas1 | Gas2 | T1D | Strok | CNS |
| RIFS | 1 | 2 | 2 | 9 | 12 | 5 | 4 | 5 |
| CFS | 68 | 39 | 115 | 300 | 237 | 166 | 138 | 39 |
| PAM | 109 | 4 | 44 | 29 | 500 | 48 | 3 | 20 |
| RRF | 9 | 3 | 7 | 16 | 9 | 28 | 14 | 29 |
| McTw | 4 | 2 | 3 | 4 | 2 | 6 | 1 | 4 |

Table 2

As shown in Table 1 and Table 2, the classification accuracy and the model complexity for the five algorithms was investigated. According to Table 1, RIFS achieves 2.3% lower than CFS in mAcc, but 3.5%, 5.9 %and 1.4 % better than PAM ,RRF and McTwo, respectively. But number of extracted features of RIFS algorithm is only 4.8 / 100 of CFS algorithm, significantly less than the CFS algorithm. Take the dataset Adenoma as an example, both RIFS and CFS achieved 1 in mAcc as shown in table 1, but RIFS used only two features compared with 39 features selected by CFS. It can be seen that the RIFS algorithm can not only extract the feature subset with better performance, but also has the advantage of a low number of features .

### Summarized performance evaluation

【重点强调，比CFS和McTwo性能类似，且优于其他算法。同时在其他算法结果较差的数据集有较大改善。所以RIFS是对其他特征选择算法的有益补充。】

RIFS performs better than the algorithms PAM, RRF and McTwo, but worse than CFS, as shown in Table 1. Take the Comparison of RIFS and CFS algorithms as an example, the mAcc calculated by algorithm RIFS in the five datasets Prostate, Colon, ALL3, Gas2, CNS were 0.951, 0.935,0.856,1, 0.883, respectively, outperforms algorithm CFS. The mAcc calculated by algorithm RIFS in the four datasets DLBCL , ALL1, Gas1, Adenoma were 0.987, 1, 0.972, 1, respectively, performs equally well with algorithm CFS. Naturally, the mAcc in the remaining datasets performs worse. Here leads to a formula CT(A,B)=(win/tie/lose) to measure the numbers of datasets that algorithm A performs better, equally well and worse compared with algorithm B by the measurement maximal accuracy mAcc. With the comparison between algorithm RIFS and algorithm CFS, CT(RIFS,CFS)=(5/4/8). The following table 3 shows the comparison between algorithm pairs from RIFS, CFS, PAM RRF , and McTwo. The column and row of RIFS are highlighted in bold.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CT(A,B) | RIFS | CFS | PAM | RRF | McTwo |
| RIFS | 0/17/0 | 5/4/8 | 12/1/4 | 15/0/2 | 9/2/6 |
| CFS | 8/4/5 | 0/17/0 | 16/1/0 | 16/1/0 | 14/2/1 |
| PAM | 4/1/12 | 0/1/16 | 0/17/0 | 13/0/4 | 4/1/12 |
| RRF | 2/0/15 | 0/1/16 | 4/0/13 | 0/17/0 | 2/0/15 |
| McTwo | 6/2/9 | 1/2/14 | 12/1/4 | 15/0/2 | 0/17/0 |

Table 3

## Conclusions

A good predictor of tumor markers has two characteristics, a relatively high prediction accuracy and a relatively small number of features. In this experiment, we selected the respective feature subset after the calculation of 17 data samples. The result combined with the prediction accuracy and the characteristic of the number of indicators shows that RFIS algorithm performs as well or better than the other existing feature selection algorithms. performance compared with other algorithms verifies the conjecture of RIFS algorithm, illustrates the effectiveness and rationality of the RIFS algorithm.

## References