# Two stones make a perfect diamond

Ruochi Zhang, 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

Hypothesis: Two stones may be fused into a perfect diamond.

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" \t "_blank), 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.

## Material and Methods【暂时保留这一节，根据需要增删】

### 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 th 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

以下每一部分均为这样的比较结果：

* 3个filter、3个wrapper特征选择算法
* 5倍交叉验证的分类算法（SVM、decision tree、Naïve Bayes）accuracy为性能指标（其他指标最好也计算保存，包括Sn, Sp, Acc, MCC）

### 【两个容易的数据集（但是其他算法也没有获得100%准确性）】

### 【两个困难的数据集】

### 【统计分析17个数据集的综合情况】

### 【methylation的数据矩阵】

* 分类：头颈癌 vs 头颈正常组织：
  + <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25093>
* 2个数据集分别是：
  + <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25089>
  + <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25083>
* 需要映射为基因数据
  + 每一个特征对应的基因组位置信息：  
    <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL8490>
  + 由于上面基于的是人类基因组版本36.1，所以需要转换为最新的人类基因组：  
    <http://www.ncbi.nlm.nih.gov/genome/tools/remap>
  + 最新版本的人类基因组可以在这里找到（版本为GRCh38.p7）：  
    <ftp://ftp.ncbi.nih.gov/genomes/Homo_sapiens/Assembled_chromosomes/>
  + 人类一个基因有多个exon组成，这个exon内部所有的methylation数据取“平均值”，作为这个exon特征。我们使用这个exon来作为最终的特征，进行“头颈癌”和“头颈正常组织”的二分类算法研究。

### 【lncRNA的数据矩阵】

* 数据来源：  
  <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55433>
* 直接下载：Series Matrix File(s)

## Conclusions

## References