Authors' Reply

Dear Editors,

We would like to thank you and the anonymous reviewers for the valuable comments and constructive suggestions. We have carefully revised the paper by addressing all comments and suggestions in the reports. The revised parts are marked in RED for easily checking in the revised version. The following is a point-by-point statement of changes and replies in response to the reviewer' comments.

Answers to Reviewer #1:

本文提出了一种特征提取方法，用于基因对之间的特征提取。该问题是机器学习相关的生物信息学的核心问题，可以应用于蛋白相互作用，基因调控关系预测等。全文结构合理，语言流畅，图表质量高。(The comments is written in Chinese)

Response: Thanks for the encouraging comments. We greatly appreciate your recommendation of accepting the manuscript.

Answers to Reviewer #2:

This work proposed a new approach to select feature genes and feature gene pairs on the binary-value gene expression data. The idea is interesting and its effectiveness is extensively verified in the experiments.

Response: Thank you for the encouraging comments. We greatly appreciate your recommendation of accepting the manuscript.

Answers to Reviewer #3:

The results in this work demonstrated that the proposed feature selection algorithm FSGGP works best on the investigated datasets. More feature selection algorithms may be incorporated in this comparative study.

Response: Thanks for your valuable comments. In the revised manuscript, we have added the linear forward selection algorithm (LFS) as a compared algorithm. The proposed algorithm and the four comparison algorithm were implemented in four datasets. The results show that the algorithm has better performance than the four comparison algorithms MIFS, MICE, LFS and FeatKNN.

Feature identification for phenotypic classification based on genes and gene pairs

Yansen Su[1], Yanxin Li[2], Zheng Zhang[3], Linqiang Pan[3,4*],
1 Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China
2 Department of Cardiovascular Internal Medicine, The Third Hospital of Xingtai, Xingtai 054000, Hebei, China
3 Key Laboratory of Image Information Processing and Intelligent Control, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
4 School of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, Henan, China
∗ To whom correspondence should be addressed. Email: lqpan@mail.hust.edu.cn.

Abstract

Background The classification of phenotypes on microarray data has drawn much attention in last few years. The known methods mainly focused on the selection or construction of features based on either genes or gene pairs on continuous-value gene expression data. However, few researches have been implemented to identify useful features based on both genes and gene pairs on binary-value gene expression data.

Results In this work, we proposed a new algorithm, called FSGGP, to select both feature genes and feature gene pairs on the binary-value gene expression data to improve two-phenotype classification. We calculated the uncertainty coefficient which represented how well a phenotype was described by a gene or gene pair under some possible relationship, and the exact relationship between the gene or gene pair and the phenotype was identified by the value of uncertainty coefficient. Furthermore, the closeness between genes or gene pairs and phenotypes was calculated, and the genes or gene pairs closely related with phenotypes were selected. The redundancy of genes and gene pairs as features was calculated by cross entropy on the binary data, and the redundant feature genes or gene pairs were eliminated. The optimal feature sets were obtained by the wrapper based forward feature selection for three classical classifiers. The algorithm was experimentally assessed on four public datasets. The results showed that algorithm FSGGP had better performance over four known feature selection algorithms based on either genes or gene pairs in terms of the average classification error rates.

Conclusions We developed an algorithm to select both feature genes and feature gene pairs on the binary-value gene expression data, where the selection of feature gene pairs was implemented by identifying the higher logical relationship between gene pairs and phenotypes. The comparison with four known feature selection algorithms suggests that feature selection algorithms based on both genes and gene pairs can achieve better performance than feature selection algorithms based on either genes or gene pairs, and the identification of higher logical relationship is an effective approach for the selection of feature gene pairs.

Keywords: Classification; Phenotype; Gene; Gene pair

1 Introduction

At a molecular level, microarrays have been widely used to obtain transcription abundance of tens of thousands of genes in various conditions [1–3]. The microarray experiments have been done on the samples of distinct phenotypes (e.g., healthy/diseased, resistant/susceptible, flowering/non-flowering) [4,5]. In some experiments, samples are labeled with phenotypic information; while, in other ones, there is no phenotypic information of a sample. It deserves to find key molecules for classification based on the gene expression data of labeled samples and apply the molecules to the samples without labels for diagnosis or classification. In general, the classification of phenotypes has drawn great attention of research community [6, 7].

A gene expression profile data is a matrix, where each row represents a gene and each column represents a sample. The elements of the gene expression profile data mainly have two forms: the binary values and the values in the interval $[0, 1]$ (e.g., the data used in [8–11]). The gene expression data with binary-values (resp., the values in the interval $[0, 1]$) is shortly referred as binary data (resp., continuous data). In a binary data, '0' represents that a gene is not expressed in a sample; while '1' means that a gene is expressed in a sample [12]. A real number in the interval $[0, 1]$ denotes the probability that a gene can be detected in a sample [13]. In a gene expression profile data, the number of genes is generally much larger than that of samples [14]. Thus, it is necessary to reduce the dimension of gene expression profile data for efficient classification.

There is a lot of work in the dimension reduction based on the continuous data (e.g., researches in [15–19]), while there are relatively few researches dedicated to reduce dimension on the binary data. The binary data

can illustrate the inactive and active states of genes [11], which could be the discretization of the continuous data without necessary information loss [10]. Compared with the continuous data, the binary data may be more easily measured and require fewer micrograms of total RNA for a comprehensive screen [20]. The efficient dimension reduction based on the binary data may improve classification accuracy with lower cost.

Feature selection methods can select useful single features, such as the mutual information based feature selection (MIFS) algorithm, the mutual information and cross entropy (MICE) algorithm, and the Linear Forward Selection (LFS) algorithm [21–23]. In some cases, feature pairs can achieve better classification performance than single features contained in the feature pairs [17,24]. However, the limited knowledge about the relationships between features leads to little biological or physical interpretations of feature pairs. Furthermore, most of the known excellent classifiers, which require inputs to be single features, are hard to use feature pairs for classification [25]. Therefore, it is necessary to select the feature pairs with exact relationships, and then transform the feature pairs into new single features for classification.

For continuous data, there are known feature transform methods, which construct new features from the original features [26] (e.g., the feature construction based on k-nearest neighbors (FeatKNN) algorithm [17]). For binary data, there is the possibility that samples with similar values of feature pairs are from different classes. In this case, the feature transform method for continuous data does not work for binary data, since the value of a new feature may have a high variance. Furthermore, the known feature transform methods always focus on feature pairs for classification. Therefore, in order to enhance classification, it is necessary to develop feature selection algorithms that take both genes and gene pairs into consideration.

In this work, we developed a new algorithm to select feature genes and feature gene pairs on binary-value gene expression data for two-phenotype classification. We initially selected the genes and gene pairs closely related with phenotypes. Then, we constructed new features based on the relationships between gene pairs and phenotypes, and eliminated the redundant new features. Finally, the feature subset with the minimum classification error was obtained by wrapper based forward feature selection. The proposed algorithm took full advantage of genes and gene pairs on binary-value gene expression data. We tested the algorithm on four public datasets. The results showed that the algorithm had better performance than the four known algorithms MIFS, MICE, LFS, and FeatKNN.

2 Materials and Methods

2.1 Dataset

The subtypes of NSCLC dataset was obtained by selecting genes from a dataset presented in [27]. The dataset presented in [27] contained the binary expression data of 40,233 genes in 210 adenocarcinoma (AC) samples and 144 squamous cell carcinoma (SCC) samples. We selected 2,765 genes from 40,233 genes whose expression data in AC samples were likely different with those in SCC samples. The samples were arranged in the order that the front 210 samples were AC samples and the last samples were SCC samples. The subtypes of NSCLC dataset contained a gene data and a phenotype profile data. The gene data was a 2765×354 matrix, where each row represented a gene and each column meant a sample. The phenotype profile data was a 2×354 matrix, where the first row represented AC and the second row meant SCC, and each column represented a sample. The detail information about the gene data and the phenotype profile data in the subtype of NSCLC dataset were presented in the Supplementary Material.

The NSCLC-normal dataset was from the binary expression data of 40,233 genes in 46 NSCLC samples and 45 normal samples [27]. We selected 5,764 genes whose expression data in NSCLC samples were likely different from those in normal samples. The obtained NSCLC-normal dataset included a 5764×91 gene data matrix and a 2×91 phenotype profile data matrix, where the first row of the phenotype profile data matrix represented NSCLC and the second one represented normal. For more information about the gene data matrix and phenotype profile data matrix please refer to the Supplementary Material.

The Arabidopsis dataset were from European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). The 47 samples in long day condition were from E-ATM-1304, E-GEOD-6906, E-MEXP-1299, and GSE11708. The 72 samples in short day condition were from E-ATM-1304, E-GEOD-6906, E-MEXP-1299 and GSE11708. Each sample contained the raw data of 22,875 probes. We converted the raw data to the binary expression data by the MAS5.0 algorithm, and selected the probes which detected single genes by the correspondence between probes and genes. Then, we selected 1,079 genes whose expression data in long day condition were likely different with those in short day condition. Finally, we obtained the binary data of 1,079 genes in 47 long day condition samples and 72 short day condition samples. The gene data of the dataset was a 1079×119 matrix, and the phenotype profile data of the dataset was a 2×117 matrix, where the first row of the phenotype profile data matrix represented Arabidopsis in long day condition and the second one represented Arabidopsis in short day condition (see the

Supplementary Material).

According to the resistance to phytophthora sojae, soybean cultivars mainly falled into two categories: plants resistant to phytophthora sojae (resistant cultivars) and those susceptible to phytophthora sojae (susceptible cultivars). The soybean dataset from GSE9687 contained the raw data of 61,170 probes in 46 resistant samples and 45 susceptible samples. We selected the genes whose expression data in resistant cultivars samples were likely different with those in susceptible cultivars samples, and we obtained the binary expression data of 3992 genes in 46 resistant samples and 45 susceptible samples. The gene data of the soybean dataset was a 3992×91 matrix. The phenotype profile data of the soybean dataset was a 2×91 matrix, where the first row represented resistant cultivars and the second row represented susceptible cultivars (see the Supplementary Material).

2.2 Calculation of closeness between features and classes

In this work, features or feature pairs closely related with classes are selected by calculating the closeness between features or feature pairs and classes. The closeness measures the extent that features or feature pairs are related with classes. The extent that a class is described by a feature or feature pair contributes to the closeness between the feature or feature pair and the class. Besides, the extent that a feature or feature pair is described by a class also contributes to the closeness. In what follows, we present the definition of the closeness between a feature and a class, and then the definition of the closeness between a feature pair and a class.

The uncertainty coefficient $U(B|f_i^l(A))$ represents how well a class B is described by a feature A under a lower logic function f_i^l. The value of $U(B|f_i^l(A))$ is calculated as follows:

$$U(B|f_i^l(A)) = \frac{H(B) + H(f_i^l(A)) - H(B, f_i^l(A)))}{H(B)}, \quad (1)$$

where $i \in \{1, 2\}$; l is the symbol for lower logic functions; $H(B)$ is the entropy of B; $H(f_i^l(A))$ is the entropy of $f_i^l(A)$; $H(B, f_i^l(A))$ is the joint entropy of B and $f_i^l(A)$.

The uncertainty coefficient for B given A, denoted by $U(B|A)$, is the maximum of $U(B|f_1^l(A))$ and $U(B|f_2^l(A))$. By the fact that $H(f_1^l(A)) = H(f_2^l(A))$ and $H(B, f_1^l(A)) = H(B, f_2^l(A))$, we have $U(B|A) = U(B|f_1^l(A)) = U(B|f_2^l(A))$. If the value of $U(B|A)$ is greater than a given threshold, then it means that there is a relationship between feature A and class B.

The confidence of $B = f_i^l(A)$, denoted by $Conf(f_i^l(A) = B)$, represents the probability of the relationship $B = f_i^l(A)$ between A and B. The value of $Conf(f_i^l(A) = B)$ is calculated as follows:

$$Conf(f_i^l(A) = B) = \frac{p_{11}}{p_{10} + p_{11}},$$

where p_{11} is the joint probability of the occurrence of $(1, 1)$ for the logic function $f_i^l(A)$; p_{10} is the joint probability of the occurrence of $(1, 0)$ for the class B.

The confidence $Conf(f_i^l(A) = B)$ is used to identify the type of the relationship between feature A and class B. Specifically, if $Conf(f_{i'}^l(A) = B)$ is the maximum of $Conf(f_1^l(A) = B)$ and $Conf(f_2^l(A) = B)$, then the exact relationship between feature A and class B is considered to be the lower logic relationship $f_{i'}^l$.

Let $f_{i'}^l$ be the exact lower logic relationship between feature A and class B, then the reverse uncertainty coefficient $U(A|B)$, representing how well the feature A is described by the class B under the function $f_{i'}^l$, is defined as follows:

$$\begin{aligned} U(A|B) &= U(f_{i'}^l(A)|B) \\ &= \frac{H(f_{i'}^l(A)) + H(B) - H(f_{i'}^l(A), B)}{H(f_{i'}^l(A))}, \end{aligned} \quad (2)$$

where $H(f_{i'}^l(A))$, $H(B)$, $H(f_{i'}^l(A), B)$ have the same meaning as those in e.q (1).

The closeness between feature A and class B, denoted by $U(A, B)$, is defined as:

$$U(A, B) = \frac{U(B|A) + U(A|B)}{2}.$$

Similarly, the uncertainty coefficient $U(C|f_j^h(A, B))$ represents the degree to which the higher logic combination $f_j^h(A, B)$ of features A and B describes class C. The value of $U(C|f_j^h(A, B))$ is $(H(C) + H(f_j^h(A, B)) - H(C, f_j^h(A, B))))/H(C)$, where h is the symbol for higher logic functions; $j \in \{1, 2, 3, 4, 5\_1, 5\_2, 6\_1, 6\_2, 7, 8\}$; $H(C)$ and $H(f_j^h(A, B))$ are the entropy of C and $f_j^h(A, B)$, respectively; $H(C, f_j^h(A, B))$ is the joint entropy of C and $f_j^h(A, B)$.

The uncertainty coefficient for class C given features A and B, denoted by $U(C|A, B)$, is the maximum of the values $U(C|f_j^h(A, B)$, $j \in \{1, 2, 3, 4, 5\_1, 5\_2, 6\_1, 6\_2, 7, 8\}$. If the value of $U(C|A, B)$ is greater than a given threshold, then it means that there is a higher logic relationship between features A, B and class C.

The confidence of $C = f_j^h(A, B)$, denoted by $Conf(f_j^h(A, B) = C)$, is calculated as follows:

$$Conf(f_j^h(A, B) = C) = \frac{p'_{11}}{p'_{10} + p'_{11}},$$

where p'_{11} represents the joint probability of occurrence of $(1, 1)$ for the higher logical relationship $f_j^h(A, B)$; p'_{10} represents the joint probability of occurrence of $(1, 0)$ for the class C. The confidence $Conf(f_j^h(A, B) = C)$ is

used to identify the type of the relationship between features A, B and class C. Specifically, if $Conf(f^h_{j'}(A,B) = C) = \max\{Conf(f^h_j(A,B) = C) \mid U(C|f^h_j(A,B)) = U(C|A,B)\}$, then the exact relationship between features A, B and class C is considered to be the higher logic relationship $f^h_{j'}$.

Let $f^h_{j'}$ be the exact higher logic relationship between features A, B and class C, then the reverse uncertainty coefficient $U(A,B|C)$, representing how well the feature pair (A,B) is described by the class C under the logical function $f^h_{j'}$ between features A, B and class C, is defined as follows:

$$\begin{aligned} U(A,B|C) &= U(f^h_{j'}(A,B)|C) \\ &= \frac{H(C) + H(f^h_{j'}(A,B)) - H(C, f^h_{j'}(A,B)))}{H(f^h_{j'}(A,B))}, \end{aligned}$$

where $j' \in \{1,2,3,4,5\_1,5\_2,6\_1,6\_2,7,8\}$; $H(C)$ and $H(f^h_{j'}(A,B))$ are the entropy of C and $f^h_{j'}(A,B)$, respectively; $H(C, f^h_{j'}(A,B))$ is the joint entropy of C and $f^h_{j'}(A,B)$.

The closeness between feature pairs (A,B) and class C is defined as follows:

$$U((A,B),C) = \frac{U(C|A,B) + U(A,B|C)}{2}.$$

2.3 Selection of features and feature pairs

In this work, we refer to features or feature pairs, which could be used to achieve the desired classification accuracy, as the effective features or feature pairs. In order to select the effective features, we take the following two hypotheses: The effective features are the features closely related with the class; If a feature is closely related with the class, then a feature pair which contains the feature is also closely related with the class.

Suppose the number of features is n, and the class is C. We obtain the feature set SS which contains m effective features and m effective feature pairs as follows (where we suppose that n is largely greater than m):

(a) Suppose $F = \{g_1, g_2, \cdots, g_n\}$, and $SS = \emptyset$;

(b) Calculate the closeness between feature g_i and class C, where $i \in \{1,2,\cdots,n\}$;

(c) Select the feature g^*_i which has the largest closeness with class C, add g^*_i into SS, and delete g^*_i from F;

(d) Calculate the closeness between feature pair (g^*_i, g_j) and class C, where g_j is a feature in F;

(e) Select the feature pair (g^*_i, g^*_j) which has the largest closeness with class C, add (g^*_i, g^*_j) into SS, and delete g^*_j from F;

(f) Repeat steps (b)-(e), until m features and m feature pairs are obtained.

2.4 Calculation of redundancy

Several methods have been proposed to measure the dependency of variables, such as mutual information, entropy, and cross entropy. In what follows, we introduce the notion of cross entropy that was used in this work to compute the redundancy of a feature set [22].

The cross entropy of $f(X)$ and $g(X)$, denoted by $D(f(X), g(X))$, measures the difference of two probability distributions $f(X)$ and $g(X)$. The value of $D(f(X), g(X))$ is calculated as:

$$D(f(X), g(X)) = \sum f(X) log \frac{f(X)}{g(X)}.$$

If $f(X)$ is equal to $p(x_1, x_2 \cdots, x_n)$ and $g(X)$ is equal to $p(x_1)p(x_2)\cdots p(x_n)$, then $D(f(X), g(X))$ is $\sum \cdots \sum p(x_1, \cdots, x_n) log[p(x_1, \cdots, x_n)/p(x_1)\cdots p(x_n)]$. For convenience, $D(f(X), g(X))$ is written as D_n for short.

If $x_1, \cdots, x_n$ are independent, then $p(x_1, \cdots, x_n)$ is equal to $p(x_1)\cdots p(x_n)$. Then, $D_n = 0$. Otherwise, D_n is $\sum \cdots \sum p(x_1, \cdots, x_n) logp(x_1, \cdots, x_n) - \sum_{i=1}^n p(x_i) logp(x_i)$. In this case, $D_n > 0$.

Let $S = -\sum \cdots \sum p(x_1, \cdots, x_n) logp(x_1, \cdots, x_n)$ and $S_i = -\sum_{i=1}^n p(x_i) logp(x_i)$, then

$$D_n = -S + \sum_{i=1}^n S_i. \tag{3}$$

Since $S_i \leq S$ $(i = 1, 2, \cdots, n)$, we have

$$S_1 + S_2 + \cdots + S_n \leq nS. \tag{4}$$

By e.q. (3) and e.q. (4), we have $D_n = S_1 + S_2 + \cdots + S_n - S \leq (n-1)S$. So, D_n can be normalized as $\bar{D}_n = (S_1 + S_2 + \cdots + S_n - S)/((n-1)S)$, where $0 \leq \bar{D}_n \leq 1$.

In general, $\bar{D}_n$ measures the dependency of n variables. The larger the value of $\bar{D}_n$ is, the more dependent the variables are. In order to select independent features, the threshold of independence should be set. If the threshold of independence is T, and a feature set has $\bar{D}_n \leq T$, then the features in this feature set are considered to be independent.

2.5 Elimination of redundant features

Let $SS = \{g_{i_1}, g_{i_2}, \cdots, g_{i_{2m}}\}$ be a feature set with features ordered by their values of closeness with a class (that is, the first element of the set represents the feature with the largest closeness value with the class, and the last one is the feature with the smallest closeness value with the class), we eliminate the redundant features and obtain the non-redundant feature set S as follows:

(a) Set $S = \{g_{i_1}\}$, $j = 2$ and the threshold of independence be T.

(b) If $j \leq 2m$, calculate $\bar{D}_n(S \cup \{g_{i_j}\})$.
If $\bar{D}_n(S \cup \{g_{i_j}\}) < T$, then $S = S \cup \{g_{i_j}\}$ and go to step (c).
If $\bar{D}_n(S \cup \{g_{i_j}\}) \geq T$, then go to step (c).
If $j = 2m + 1$, stop.

(c) Set $j = j + 1$, go to step (b).

2.6 Matching rate

The matching rate of a gene or gene pair is defined as follows:

$$M(g) = \frac{n}{N},$$

where g represents a gene or gene pair; n represents the number of samples in which the value of a gene or gene pair under a logic function is equal to the value of a phenotype (the value of a phenotype in a sample is either 0 or 1, meaning the absence or presence of a phenotype); N represents the number of all samples.

The matching rate of a gene or gene pair is used to evaluate the ability of a gene or gene pair for classification. Suppose X is a feature set, then the mean value of the matching rate of all elements in X is the matching rate of X, denoted by $Ratio(X)$.

2.7 Classifier

In this work, the following three classifiers were used for the comparison of the classification performance: Naive Bayes, diagonal linear discriminant (DLD) and linear discriminant analysis (LDA), since it was reported that they were among the most efficient classifiers for microarray data classification [28, 29]. In what follows, we briefly introduce these three classifiers.

Naive Bayes is based on the Bayesian theory. Suppose there are n features $X = \{x_1, x_2 \cdots, x_n\}$, as well as two classes $C = \{C_1, C_2\}$. According to the Bayesian theory, the probability of a sample belongs to C_i, $i \in \{1, 2\}$, is $p(C_i|X) = \frac{p(X|C_i)p(C_i)}{p(X)}$. Suppose the attributes are independent, then $p(X|C_i) = \prod_{k=1}^{n} p(x_k|C_i)$. If $p(C_i|X) \geq p(C_j|X)$, then a sample is classified into the class C_i, where $j \in \{1, 2\}$ and $j \neq i$.

The classifier DLD turns out the minor variant of samples. Suppose the gene expression data is $x = (x_1, x_2 \cdots, x_n)$, and the class number is either 1 or 2. The DLD axis a is computed as follows:

$$a = M^{-1}(\mu_1 - \mu_2), \tag{5}$$

where μ_1 and μ_2 are the mean values of the gene expression data in C_1 and C_2, respectively; M is a diagonal variance matrix, and its element (i, i) is $\sigma_i^2 = \frac{(n_1-1)\sigma_{1,i}^2+(n_2-1)\sigma_{2,i}^2}{(n_1+n_2-2)}$, $\sigma_{t,i}$ is the standard deviation of gene i in C_t, n_t is the number of samples in C_t, $t \in \{1, 2\}$.

The classifier LDA aims to find a linear projection matrix based on the training data, to make the ratio of the variance between the classes to the variance within the classes as large as possible. The work flow of LDA is as follows: establish the linear discriminant function based on the training data, and determine the class number of a sample by the linear discriminant function.

2.8 Error estimation

In order to evaluate the classification performance of a feature set, the classification error rate of a feature set are calculated. Cross-validation is the most commonly used error estimate method [35, 36]. However, it was pointed that this method is not the best method for classification on the small-sample microarray data [30]. That is because the randomly selected samples lead to the large difference of error estimation of a feature set for classification. The 0.632 bootstrap estimator could help to solve this problem. So, in this work, we use the 0.632 bootstrap estimator for error estimation. In what follows, the 0.632 bootstrap estimator is briefly introduced.

We randomly select samples from original samples with replacement for n_0 times, and obtain n_0 training samples. Several of these training samples may be the same. The original samples which are not included in training samples are testing samples. The probability that an original sample is a training sample is $(1 - \frac{1}{n_0})^{n_0} \approx e^{-1} \approx 0.368$. In other words, the training set contains about 63.2% original samples and the testing set contains about 36.8% original samples. The estimation of the error rate of a feature set, denoted by err, is calculated as follows: $err = 0.632e_1 + 0.368e_2$, where e_1 represents the error rate estimation of the feature subset in the training set, and e_2 denotes the error rate estimation of the feature subset in the testing set.

2.9 Selection of the optimal feature set

In order to improve classification performance, the optimal feature subset deserves to be chosen from the non-redundant features which are closely related with classes. The filter methods and the wrapper methods are two categories of feature selection methods. The filter methods need less computation, but they could not select the most relevant feature set for classifiers. The wrapper methods could find the better features for classifiers, but they need more computational time [22]. In this work, we use a wrapper algorithm to select an optimal feature set from non-redundant features.

Suppose there are p non-redundant features $\{g_1, \cdots, g_p\}$ which are closely related with the class

C, and a given classifier M. The algorithm to select the optimal feature set from these features is as follows:

(a) Let $S = \{g_1, \cdots, g_p\}$, and $R = \emptyset$;

(b) The error rate estimation err_i for set $\{g_i\}$ is calculated by 0.632 bootstrap estimator, $i = 1, \cdots, p$;

(c) If $err_{i^*} = \min\{err_1, err_2, \cdots, err_p\}$, then $R = R \cup \{g_{i^*}\}$, $S = S - \{g_{i^*}\}$, and $min\_err = err_{i^*}$;

(d) For $g_j \in S$, the error rate estimation err_j for $R \cup \{g_j\}$ is calculated by 0.632 bootstrap estimator;

(e) If $err_{j^*} = \min\{err_j | g_j \in S\}$, then $new\_min\_err = err_{j^*}$;

(f) If $new\_min\_err < min\_err$, then $min\_err = new\_min\_err$, $R = R \cup \{g_{j^*}\}$, and $S = S - \{g_{j^*}\}$; If $new\_min\_err \geq min\_err$, then $S = S - \{g_{j^*}\}$;

(g) Repeat steps (d)-(f), until $min\_err = 0$ or $S = \emptyset$.

3 Results and discussion

We proposed a feature selection algorithm based on genes and gene pairs (FSGGP) for two-phenotype classification on binary-value gene expression data. The algorithm FSGGP has three stages. In the first stage, we select the genes and gene pairs which are closely related with the phenotypes by the values of closeness between genes or gene pairs and phenotypes. In the second stage, the redundant features are eliminated by calculating the cross entropy. In the third stage, the feature subset with the minimum classification error is obtained by the wrapper based forward feature selection. To test the efficacy of the proposed method, we designed comparison experiments on four public datasets: the subtypes of non-small cell lung cancer (NSCLC) dataset, the NSCLC-normal dataset, the Arabidopsis dataset and the soybean dataset. The results of comparison showed that the algorithm FSGGP is effective for feature selection.

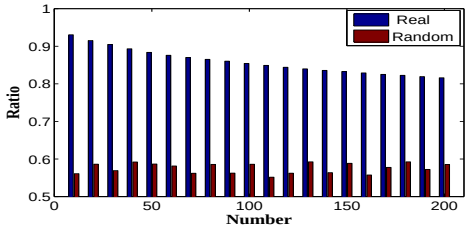
3.1 Differentially expressed genes

The following four datasets were used in this work: the subtypes of NSCLC dataset, the NSCLC-normal dataset, the Arabidopsis dataset and the soybean dataset. If the proportions of '1' in the binary data of a gene in different classes have a large difference, then the expression data of the gene in one class is considered to be different with that in another class, and the gene is assumed to be good for classification, where a class means a phenotype. In this work, if the difference between proportions is greater than 0.2, then the gene is selected for further analysis. The reasonableness of the setting of the parameter will be explained in the 'Efficiency of parameters' subsection. In this way, all genes in each dataset are classified into two categories: non-differentially expressed genes and differentially expressed genes.

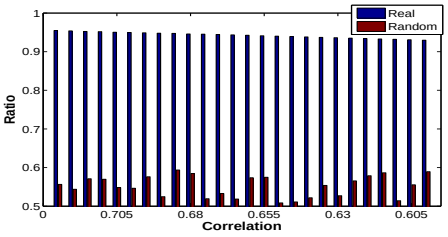
For each dataset, we randomly selected 100 genes to obtain a gene set for 1000 times, and calculated the average matching rate of the 1000 gene sets (for the notion of matching rate). We found that the non-differentially expressed genes had less ability for classification. Thus, it is reasonable to select differentially expressed genes for classification.

3.2 Identification of feature genes and feature gene pairs

We identified feature genes and feature gene pairs based on the following two assumptions: the effective features or feature pairs are closely related with the classes; if a feature is closely related with a class, then a feature pair that contains the feature is also closely related with the class. In what follows, we show the validity of these assumptions.



(a) Genes

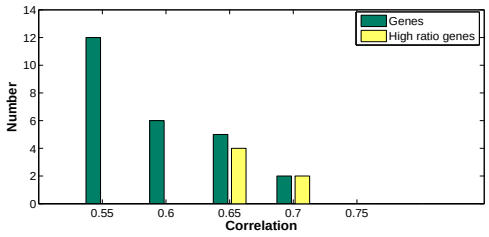


(b) Gene pairs

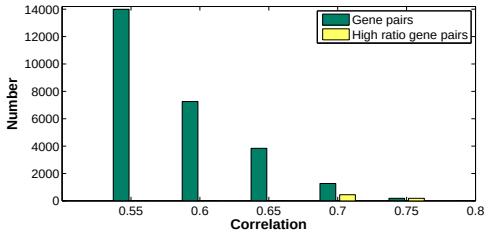
Figure 1. The correlation between the ability of a gene (gene pair) set for classification and the number of genes (gene pairs). 'Ratio' denotes the matching rate of a gene (gene pair) set. We randomly select the same number of genes (gene pairs) with that in each 'Real' set to obtain a 'Random' set.

We ranked the genes and the gene pairs in the descending order, respectively, based on their values of closeness with the classes (for the notion of closeness and the calculation of closeness values. That is, the gene with the highest closeness value had rank value 1, and the one with the lowest closeness value had the lowest rank. Similarly, we computed the closeness values between gene pairs and the classes, and then defined the ranking of gene pairs based on the closeness values. We selected

the top $10i$ genes and the top $10i$ gene pairs to form a feature set, where $i = 1, 2, \cdots, 20$. In this way, we got 20 feature sets, and calculated the matching rates of these 20 feature sets. From the results on the subtypes of NSCLC dataset, we found that the matching rates of feature sets decreased with the increase of the numbers of elements in feature sets (Fig.1(a) and Fig.1(b)). The increase of the numbers of elements in feature sets meant the decrease of the average value of closeness of the elements in the feature sets, since the genes or gene pairs were ranked by the values of closeness. Besides, the matching rates of the selected feature sets were much greater than those of randomly selected features (Fig.1(a) and Fig.1(b)). We further analyzed the correlation between the matching rates of feature sets and the numbers of elements of feature sets on the other three datasets: the NSCLC-normal dataset, the Arabidopsis dataset and the soybean dataset, and obtained similar results. In general, all results showed that the higher the closeness of genes or gene pairs with the classes is, the higher the ability of the genes or gene pairs for classification is, which suggested that it is reasonable to select genes or gene pairs by the values of closeness for classification.



(a) Closeness between genes and phenotypes



(b) Closeness between gene pairs and phenotypes

Figure 2. The distributions of closeness. 'High ratio genes' represents the distributions of the closeness between the genes with great ability for classification and phenotypes. 'Genes' represents the distribution of the closeness between all genes and phenotypes. 'High ratio gene pairs' represents the distributions between the closeness of the gene pairs with great ability for classification and phenotypes. 'Gene pairs' represents the distribution of the closeness between all gene pairs and phenotypes.

For the subtypes of NSCLC data, the largest matching rate of genes and gene pairs were 0.930 and 0.955, respectively. We selected the genes and gene pairs with matching rate greater than 0.920 and 0.940, respectively. The distribution of the values of the closeness of these genes and gene pairs with classes were shown in Fig.2(a) and Fig.2(b). From the figures, we found that the gene with high matching rate was more likely to have high closeness value. Similar results were also observed on the other three datasets: the NSCLC-normal dataset, the Arabidopsis dataset and the soybean dataset. The above analysis showed that the effective features or feature pairs were closely related with the classes, which further suggested that it is reasonable to select genes or gene pairs by the values of closeness for classification.

In what follows, we validate another assumption: if a feature is closely related with a class, then a feature pair that contains the feature is also closely related with the class. We selected the gene pairs whose closeness values were greater than 0.7, and considered the rank values of the genes contained in these gene pairs. For the subtypes of NSCLC dataset, there were 44 gene pairs with the closeness values greater than 0.7.

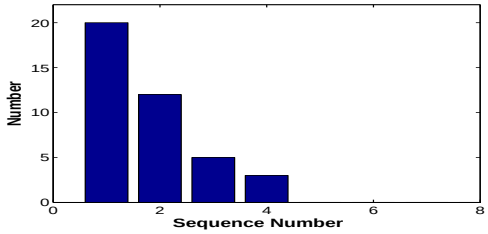


Figure 3. The distributions of the values of rank about genes. The genes with large closeness with phenotypes are used to construct gene pairs. For a gene pair, the smaller one of the rank values of the two genes contained in the gene pair are taken into consideration.

The distribution of the rank values of the genes contained in these 44 gene pairs was shown in Fig.3. The smallest rank value was 1, and the largest one was 4. Note that if a gene had smaller value of rank, then the closeness between the gene and the class was higher. So, the distribution shown in Fig.3 suggested that the gene pair with large closeness value contained a gene closely related with the class, which showed that it is reasonable to assume that if a feature was closely related with a class, then a feature pair containing the feature was also closely related with the class.

3.3 Selection of the optimal gene set

In order to check whether the selected feature sets are over fitted, for each of the above mentioned datasets, we selected the optimal feature sets for classifiers: Naive Bayes, diagonal linear discriminant (DLD) and linear discriminant analysis (LDA), respectively.

For the subtypes of NSCLC dataset, the optimal

feature set for classifier Naive Bayes contained seven elements, which were singles genes *JAK1*, *CLEC4D*, *SEC16B* and *RBM14*, and gene pairs (*SLC18A2*, *IFNE*), (*FAM62B*, *FLJ31222*) and (*TAS2R39*, *NDUFS4*). The optimal feature set for classifier DLD contained ten elements, which were seven single genes *ATOH7*, *C7orf45*, *C1orf211*, *FAM160B1*, *ABCC13*, *SEC16B* and *DIP2A*, and three gene pairs (*SLC18A2*, *IFNE*), (*FAM62B*, *FLJ31222*) and (*TAS2R39*, *NDUFS4*). It was found that the gene pairs included in the optimal feature set for DLD were the same with those for Naive Bayes. The optimal feature set for classifier LDA contained seven elements, which were two single genes *TOR1A* and *MBD1*, and five gene pairs (*SLC18A2*, *IFNE*), (*FAM62B*, *FLJ31222*), (*TAS2R39*, *NDUFS4*), (*SETMAR*, *KLHL29*) and (*ST3GAL3*, *C7orf34*). We found that (*SLC18A2*, *IFNE*), (*FAM62B*, *FLJ31222*) and (*TAS2R39*, *NDUFS4*) were also included in the feature set for DLD, which suggested that gene pairs (*SLC18A2*, *IFNE*), (*FAM62B*, *FLJ31222*) and (*TAS2R39*, *NDUFS4*) were important for classification of subtypes of NSCLC. Although few of the above genes have been confirmed to be related with the subtypes of NSCLC, these gene pairs displayed different roles in different subtypes of NSCLC. The mean error rates of the obtained optimal feature sets for Naive Bayes, DLD and LDA were 0.0196, 0.0180 and 0.0211, respectively. Thus, the classifier DLD had best performance for the classification of the subtypes of NSCLC, with the corresponding feature genes *ATOH7*, *C7orf45*, *C1orf211*, *FAM160B1*, *ABCC13*, *SEC16B*, *DIP2A*, *SLC18A2*, *IFNE*, *FAM62B*, *FLJ31222*, *TAS2R39*, *NDUFS4*.

For the NSCLC-normal dataset, the optimal feature sets for Naive Bayes, DLD and LDA were same, which contained only one gene *MAD2L1*. The gene *MAD2L1* is necessary for progression through the cell cycle [33], and this gene has been proposed to be one of the potential biomarkers for NSCLC [34]. Our result supported that the gene *MAD2L1* was a potential biomarker for NSCLC. Furthermore, all of the three classifiers Naive Bayes, DLD and LDA, with the gene *MAD2L1* as the feature set, had good performance for the classification of NSCLC and normal samples.

For the Arabidopsis dataset, classifier LDA had the minimum classification error, with the corresponding feature set containing genes *AT1G74010*, *MRN*, *KNAT1*, *ATHB33*, *AT5G16980*, *AGL42*, *AT1G60270*, *AT1G10640*, and gene pairs (*MBP1*, *AT3G51720*) and (*AT4G37970*, *WAK1*).

For the soybean dataset, classifier Naive Bayes had the minimum classification error rate, with the corresponding feature set containing genes *LOC100305826*, *LOC100305528*, *LOC100305844*, *LOC547584* and *LOC100305819*. All of these genes encode uncharacterized proteins, which suggested that these genes were good candidates for the identification of the resistant and susceptible cultivars.

Table 1. The significance of the selected feature sets on four datasets

Data	Classifier	Error rate	
		Selected feature	Mean(std)
1	DLD	0.047 ± 0.005	0.048 ± 0.005
2	Bayes	0	0
3	LDA	0.081 ± 0.023	0.145 ± 0.056
4	Bayes	0.147 ± 0.018	0.289 ± 0.059

'1' represents the subtypes of NSCLC dataset. '2' represents the NSCLC-normal dataset. '3' represents the Arabidopsis dataset. '4' represents the soybean dataset. 'Selected feature' represents the error rate for the selected feature set. 'Mean(std)' represents the average error rate for 1000 runs.

We carried out the 0.632 bootstrap estimator to check the significance of the selected feature set for each dataset. We repeated the experiments for 1000 times. The average error rates and the standard deviations were shown in Table 1. We can observe that the selected feature sets were not over fitted.

3.4 Efficiency of parameters

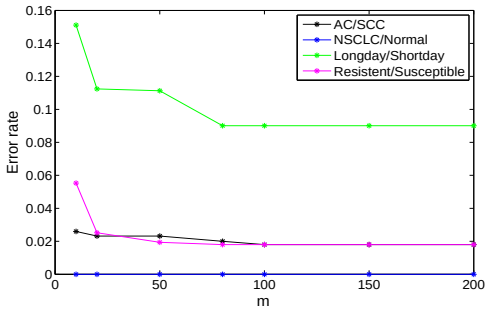
The parameters considered in this work included the ratio for selecting different expressed genes in the data processing, the number of feature genes and gene pairs m, the threshold of independence T for eliminating redundant feature genes and gene pairs.

In the data processing, we set the ratio for selecting different expressed genes to be 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, and examined the corresponding classification error rates, respectively. The classifications had the same error rate when the ratio was 0.05, 0.10, 0.15 and 0.20. When the ratio was greater than 0.2, the classification error rate increased. So, it is reasonable to set the ratio to be 0.2.

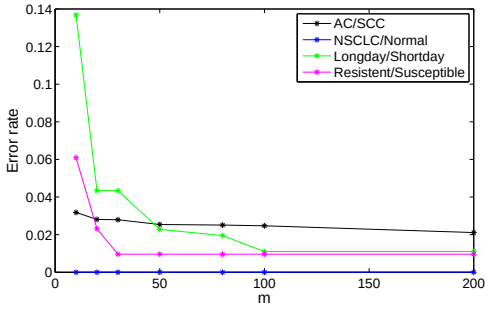
In the process of selecting feature genes and gene pairs, different values of m led to different feature sets. In what follows, the relationship between the values of m and the error rates was discussed. If the value of m was small, then the error rate may be high. On the other hand, if the value of m was large, then the complexity of computation increased. So, in our experiments, we take $m = 10, 20, 30, 50, 80, 100, 150, 200$.

For classifier Naive Bayse, the classification on the subtypes of NSCLC dataset had the minimum error rate when m was 100. We also examined the relationship between the values of m and the error rates on the NSCLC-normal dataset, and found that all classifications with different values of m had the minimum error rate zero. When m was greater than 100, the minimum error rate decreased slowly with the increase of the value of m on the Arabidopsis dataset. On the soybean dataset, the classification had minimum error rate when m was near-

ly 100.



(a) DLD



(b) LDA

Figure 4. The correlation of the error rates and the values of m for classifier DLD and LDA. We rank the genes and gene pairs in the descending order by the values of closeness, respectively. We select the top m genes and the top m gene pairs to form a feature set, where $m = 10, 20, 30, 50, 80, 100, 150, 200$. 'AC/SCC' represents the subtypes of NSCLC dataset. 'NSCLC/Normal' represents the NSCLC-normal dataset. 'Longday/Shortday' represents the Arabidopsis dataset. 'Resistant/Susceptible' represents the soybean dataset. 'Error rate' denotes the classification error rate obtained by 0.632 bootstrap estimator.

For classifiers DLD and LDA, the error rate decreased with the increase of the value of m. In general, on the four datasets, the classifications had the minimum error rate when $m = 100$ (Fig.4(a) and Fig.4(b)). So, we take $m = 100$ in the process of selecting feature genes and gene pairs.

In our experiment, the value of T was varied from 0.1 to 0.9 with step size 0.1, and it was observed that the non-redundant sets were the same for the values between 0.4 and 0.9. So, we take $T = 0.5$ in the process of eliminating redundant feature genes and gene pairs.

3.5 Performance analysis

We compared the algorithm FSGGP with four known algorithms (MIFS, MICE, LFS, and FeatKNN) on the four datasets: the subtypes of NSCLC dataset, the NSCLC-normal dataset, the Arabidopsis dataset, the soybean dataset.

(1) MIFS algorithm utilizes the maximization of mutual information between features and classes to select useful features for classification [21]. (2) MICE algorithm is based on mutual information [22]. This algorithm is a two stage algorithm. In the first stage, the gene set, whose elements are single genes, is created by the mutual information and the cross entropy. In the second stage, a forward feature selection is used to find the feature subset that minimizes classification error. (3) LFS algorithm is derived from Sequential Forward Selection, and it reduces the number of attributes expansion in each forward selection step [23]. (4) FeatKNN algorithm selects the most informative gene pairs by a heuristic search [17]. For each selected gene pairs, the new feature is constructed based on the k-nearest neighbors algorithm. Note that the algorithms MIFS, MICE, and LFS select only single genes for classification, and the algorithm FeatKNN selects only gene pairs for classification.

Table 2. Classification error rates for feature sets selected by different algorithms

	Algorithm	Classification error rate		
		Bayes	DLD	LDA
1	MIFS	0.047 ± 0.004	0.048 ± 0.004	0.040 ± 0.016
	MICE	0.055 ± 0.007	0.303 ± 0.013	0.043 ± 0.013
	FeatKNN	0.110 ± 0.006	0.406 ± 0.008	0.041 ± 0.008
	LFS	0.030 ± 0.004	0.190 ± 0.063	0.037 ± 0.008
	FSGGP	0.019 ± 0.003	0.018 ± 0.006	0.021 ± 0.003
2	MIFS	0	0	0
	MICE	0	0	0
	FeatKNN	0	0.482 ± 0.010	0.060 ± 0.015
	LFS	0	0	0
	FSGGP	0	0	0
3	MIFS	0.080 ± 0.023	0.145 ± 0.055	0.056 ± 0.027
	MICE	0.090 ± 0.000	0.443 ± 0.001	0.029 ± 0.004
	FeatKNN	0.107 ± 0.012	0.394 ± 0.013	0.049 ± 0.019
	LFS	0.058 ± 0.029	0.304 ± 0.037	0.028 ± 0.016
	FSGGP	0.057 ± 0.016	0.090 ± 0.013	0.019 ± 0.018
4	MIFS	0.146 ± 0.017	0.289 ± 0.058	0.157 ± 0.025
	MICE	0.057 ± 0.035	0.175 ± 0.001	0.034 ± 0.010
	FeatKNN	0.116 ± 0.015	0.487 ± 0.009	0.119 ± 0.002
	LFS	0.107 ± 0.024	0.045 ± 0.015	0.039 ± 0.012
	FSGGP	0.001 ± 0.004	0.018 ± 0.003	0.011 ± 0.019

'1' represents the subtypes of NSCLC dataset. '2' represents the NSCLC-normal dataset. '3' represents the Arabidopsis dataset. '4' represents the soybean dataset.

In our experiment, we used 0.632 bootstrap estimator for the error estimation. Table 2 showed the classification error rates of algorithms MIFS, MICE, LFS, FeatKNN and FSGGP on the subtypes of NSCLC dataset, the NSCLC-normal dataset, the Arabidopsis dataset and the soybean dataset. For all of the four datasets, the error rates of FSGGP were about 0.0197, that is, the classification accuracy was 98.03% on average, which meant that the performance of FSGGP was better than the other four known algorithms. This result suggested that classification algorithms based on both genes and gene pairs could achieve better performance than those based on either single genes or gene pairs.

We compared the time complexity of MIFS, MICE, LFS, FeatKNN and FSGGP. Suppose that the number of genes is n, and the number of genes or gene pairs selected for classification is p. The time complexity of MIFS is $O(n)$, and that of FeatKNN is $O(pn)$. The time complexity of LFS is $O(pn)+O(p^2)$. The time complexity for MICE to have the non-redundant features is $O(n^2)$, and p features are found with a complexity $O(p^2)$. The time complexity of FSGGP to select genes and gene pairs is $O(mn)$, where m was the number of obtained features, the redundant genes and gene pairs were eliminated with the complexity $O(m^2)$, and the optimal feature set was obtained with the complexity $O(p^2)$. So, the time complexity of FSGGP was $O(mn) + O(m^2) + O(p^2)$. Since the number of genes n was generally far greater than the number of selected feature genes and gene pairs m, the time complexity of FSGGP was grater than that of MIFS and FeatKNN, and smaller than that of MICE. That is, the algorithm FSGGP had a moderate time complexity compared with the four known algorithms.

4 Conclusions

In this work, we have proposed a feature selection algorithm FSGGP for two-phenotype classification on the binary-value gene expression data. The main idea of the proposed algorithm is that the identification of the exact relationships between genes or gene pairs and phenotypes may improve classification performance, and the selection of both genes and gene pairs as features may also improve classification performance. The algorithm has three stages: initially select the genes and gene pairs closely related with phenotypes; then eliminate redundant genes or gene pairs by computing the cross entropy; finally obtain the optimal feature set for a classifier by the wrapper based forward feature selection. The proposed algorithm has the moderate time complexity compared with the four known algorithms MICE, MIFS, LFS and FeatKNN. We compared the performance of algorithms FSGGP, MICE, MIFS, LFS and FeatKNN in terms of classification error rate on the four public datasets: the subtypes of NSCLC dataset, the NSCLC-normal dataset, the Arabidopsis dataset and the soybean dataset. The obtained result showed that the performance of FSGGP was better than the other four known algorithms.

In this work, the proposed algorithm FSGGP is limited to two-class classification. Of course, the number of phenotypes for classification may be greater than two, and the algorithm FSGGP can be modified for multi-class classification.

List of abbreviations

The abbreviations are listed according to their appearance in the paper.

- FSGGP—Feature selection algorithm based on genes and gene pairs;
- The binary data—The binary-value gene expression data;
- The continuous data—The gene continuous expression data;
- MIFS—The mutual information based feature selection algorithm;
- MICE—The mutual information and cross entropy algorithm;
- LFS—The linear forward selection algorithm;
- FeatKNN—The feature construction based on k-nearest neighbors algorithm;
- NSCLC—Non-small cell lung cancer;
- AC—Adenocarcinoma;
- SCC—Squamous cell carcinoma;
- EBI—European bioinformatics institute;
- NCBI—The national center for biotechnology information;
- DLD—Diagonal linear discriminant;
- LDA—Linear discriminant analysis.

Conflict of interest

The authors declare of no conflict of interest.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (61502004, 91530320, 61672248, 91130034, 61672033), and the Innovation Scientists and Technicians Troop Construction Projects of Henan Province (154200510012). A series of suggestions made by the anonymous referees are gratefully acknowledged.

Supplementary Material

Supplementary material is available on the publishers web site along with the published article.

References

[1] Lin Jingmei, Cao Qi, Zhang Jianjun, Li Yong, Shen Bo, Zhao Zijin, Chinnaiyan Arul M, Bronner Mary P. MicroRNA expression patterns in indeterminate inflammatory bowel disease. Modern Pathol 2013;26(1):148–154.

[2] Hehe Wang, LaChelle Waller, Sucheta Tripathy, Steven K. St. Martin, Lecong Zhou, Konstantinos Krampis, Dominic M. Tucker, et al. Analysis of genes underlying soybean quantitative trait loci conferring partial resistance to Phytophthora sojae. Plant Genome 2010;3(1):23–40.

[3] Su Yansen and Meng Dazhi and Li Eryan and Wang Shudong. Analysis of gene networks for Arabidopsis flowering. Tsinghua Science and Technology 2012;17(6):682–690.

[4] Bechara Elias G, Sebestyén Endre, Bernardis Isabella, Eyras Eduardo, Valcárcel, Juan. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell 2013;52(5):720–733.

[5] Antje Hascher, Ann-Kristin Haase, Katja Hebestreit, Christian Rohde, Hans-Ulrich Klein, Maria Rius, Dominik Jungen, Anika Witten , et al. DNA Methyltransferase inhibition reverses epigenetically embedded phenotypes in lung cancer preferentially affecting Polycomb target genes. Clin Cancer Res 2014;20(4):814–826.

[6] Warnat Patrick, Eils Roland, Brors Benedikt. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC Bioinformatics 2005;6(1):265.

[7] Bush Andrew. Classification of phenotypes. Pediatr Pulm 2004;37(Suppl 26):30–33.

[8] Mo Qianxing, Wang Sijian, Seshan Venkatraman E, Olshen Adam B, Schultz Nikolaus, Sander Chris, Powers R Scott, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. P Natl Acad Sci USA 2013;110(11):4245–4250.

[9] Berestovsky Natalie, Nakhleh Luay. An evaluation of methods for inferring Boolean networks from time-series data. Plos One 2013;8(6):e66031.

[10] Zhang Zhongyuan, Li Tao, Ding Chris, Ren Xianwen, Zhang Xiangsun. Binary matrix factorization for analyzing gene expression data. Data Min Knowl Disc 2010;20(1):28–52.

[11] Bose Indrani, Ghosh Sayantari. Origins of binary gene expression in post-transcriptional regulation by microRNAs. Eur Phys J E 2012;35(10):1–8.

[12] Datta Aniruddha, Choudhary Ashish, Bittner Michael L, Dougherty Edward R. External control in Markovian genetic regulatory networks: the imperfect information case. Bioinformatics 2004;20(6):924–930.

[13] Shmueli Orit, Horn Saban Shirley, Chalifa Caspi Vered, Shmoish Michael, Ophir Ron, Benjamin Rodrig, et al. GeneNote: whole genome expression profiles in normal human tissues. CR Biol 2003;326(10):1067–1072.

[14] Chee Mark, Yang Robert, Hubbell Earl, Berno Anthony, Huang Xiaohua C, Stern David, Winkler Jim, Lockhart David J, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274(5297):610–614.

[15] Swiniarski Roman W, Skowron Andrzej. Rough set methods in feature selection and recognition. Pattern Recogn Lett 2003;24(6):833–849.

[16] Guyon Isabelle, Gunn Steve, Nikravesh Masoud, Zadeh L. Feature extraction. Foundations and Applications 2006;24(6):833–849.

[17] Hanczar Blaise, Zucker Jean-Daniel, Henegar Corneliu, Saitta Lorenza. Feature construction from synergic pairs to improve microarray-based classification. Bioinformatics 2007;23(21):2866–2872.

[18] Guyon Isabelle, Elisseeff André. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–1182.

[19] Chebrolu Srilatha, Abraham Ajith, Thomas Johnson P. Feature deduction and ensemble design of intrusion detection systems. Compute Secur 2005;24(4):295–307.

[20] Cho Yongjig, Meade Jonathan D, Shester Blake R, Walden Jamie C, Guo Zhen, Liang Peng. Proof-reading signal accuracy of gene expression by binary differential display. Biotechnol Letters 2010;32(8):1039–1044.

[21] Battiti Roberto. Using mutual information for selecting features in supervised neural net learning. IEEE T Neural Networ 1994;5(4):537–550.

[22] Bala Rajni, Agrawal RK. Mutual information and cross entropy framework to determine relevant gene subset for cancer classification. Informatica 2011;35(3):375–382.

[23] Gutlein M, Frank E, Hall M, Karwath A. Large-scale attribute selection using wrappers. IEEE Symposium on Computational Intelligence and Data Mining 2009:332–339.

[24] Bo Trond, Jonassen Ing. New feature subset selection procedures for classification of expression profiles. Genome Biol 2002;3(4):1–17.

[25] Bounhas Myriam, Ghasemi Hamed Mohammad, Prade Henri, Serrurier Mathieu, Mellouli Khaled. Naive possibilistic classifiers for imprecise or uncertain numerical data. Fuzzy Sets and Systems 2014;239:137–156.

[26] Torkkola Kari. Feature extraction by non parametric mutual information maximization. J Mach Learn Res 2003;3:1415–1438.

[27] Yansen Su, Linqiang Pan. Identification of logic relationships between genes and subtypes of non-small cell lung cancer. Plos One 2014;9(4):e94644.

[28] Yeung Ka Yee, Bumgarner Roger E, Raftery Adrian E. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 2005;21(10):2394–2402.

[29] Dudoit Sandrine, Fridly Jane, Speed Terence P. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97(457):77–87.

[30] Braga-Neto Ulisses M, Dougherty Edward R. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004;20(3):374–380.

[31] Dhanoa Bajinder S, Cogliati Tiziana, Satish Akhila G, Bruford Elspeth A, Friedman James S. Update on the Kelch-like (KLHL) gene family. Human Genomics 2013;7(1):13.

[32] Plate Markus, Li Ting, Wang Yu, Mo Xiaoning, Zhang Yingmei, Ma Dalong, Han Wenling. Identification and characterization of CMTM4, a novel gene with inhibitory effects on HeLa cell growth through Inducing G2/M phase accumulation. Mol Cells 2010;29(4):355–361.

[33] Ross Douglas T, Scherf Uwe, Eisen Michael B, Perou Charles M, Rees Christian, Spellman Paul, Iyer Vishwanath, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000;24(3):227–235.

[34] Wang Chun I, Wang Chih Liang, Wang Chih Wei, Chen Chi De, Wu Chih Ching, Liang Ying, Tsai Ying Huang, et al. Importin subunit alpha-2 is identified as a potential biomarker for non-small cell lung cancer by integration of the cancer cell secretome and tissue transcriptome. Int J Cancer 2011;128(10):2364–2372.

[35] Magis Andrew T, Price Nathan D. The top-scoring 'N' algorithm: a generalized relative expression classification method from small numbers of biomolecules. BMC Bioinformatics 2012;13(1):227.

[36] Kopriva Ivica, Filipović Marko. A mixture model with a reference-based automatic selection of components for disease classification from protein and/or gene expression levels. BMC Bioinformatics 2011;12(1):496.

[37] Tan Pang Ning, Kumar Vipin, Srivastava Jaideep. Selecting the right interestingness measure for association patterns. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002;32–41.

[38] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–537.

[39] Alon Uri, Barkai Naama, Notterman Daniel A, Gish Kurt, Ybarra Suzanne, Mack Daniel, Levine Arnold J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. P Natl Acad Sci USA 1999;96(12):6745–6750.