Zeng and Li *BMC Medical Genomics* 2014, **7**(Suppl 2):S5
http://www.biomedcentral.com/1755-8794/7/S2/S5

BMC Medical Genomics

RESEARCH Open Access

Supervised redundant feature detection for tumor classification

Xue-Qiang Zeng[1,3†], Guo-Zheng Li[2*]

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013)
Shanghai, China. 18-21 December 2013

Abstract

Background: As a high dimensional problem, analysis of microarray data sets is a challenging task, where many weakly relevant or redundant features affect overall performance of classifiers.

Methods: The previous works used redundant feature detection methods to select discriminative compact gene set, which only considered the relationship among features, not the redundancy of classification ability among features. This study propose a novel algorithm named RESI (Redundant fEature Selection depending on Instance), which considers label information in the measure of feature subset redundancy.

Results: Experimental results on benchmark data sets show that RESI performs better than the previous state-of-the-art algorithms on redundant feature selection methods like mRMR.

Conclusions: We propose an effective supervised redundant feature detection method for tumor classification.

Background

Rapid advances in gene expression microarray technology enable simultaneous measurement of the expression levels for thousands or tens of thousands of genes in a single experiment. Analysis of microarray data presents unprecedented opportunities and challenges for data mining in areas such as gene clustering, class discovery, and tumor classification [1]. A typical data set may contain thousands of genes but only a small number of samples (often less than a hundred). The number of samples is likely to remain small at least for the near future due to the expense of collecting microarray samples [2]. The nature of relatively high dimensionality but small sample size in microarray data causes the known problem of "curse of dimensionality". Therefore, selecting a small number of discriminative ones from thousands of genes is essential for successful sample classification [3-5].

Feature selection, a process of choosing a subset of features from the original, is frequently used as a preprocessing technique in data mining. It has been proved effective in reducing dimensionality, improving mining efficiency, increasing mining accuracy, and enhancing result comprehensibility [6-9]. The goal of the feature selection algorithm is to select the minimum set of features that are strongly related to the desired decision variable and have the least redundancy among them [10,11]. Existing efficient feature selection algorithms (feature ranking methods) usually assume feature independence, and assign a discriminative score to each feature. Because the interactions and correlations among features are not considered, these algorithms fail to remove redundant features completely.

Hall [12] pointed out that the prediction capability and the inter-correlation of feature subset are two important aspects in feature selection. However, the optimal feature selection requires an exponentially large search space ($O(2^m)$, where m is the number of features) [10]. In practise, researchers often resort to various approximations to determine the optimal feature subset. The prediction capability is usually estimated by the interrelation of individual feature with the target. For the issue of redundancy, many researchers attempt to explicitly address feature interactions by finding some low-order interactions, *i.e.*

† Contributed equally

[2]Department of Control Science & Engineering, Tongji University, 4800 Cao An Road, Shanghai, 201804, China
Full list of author information is available at the end of the article

2-way (one feature and the label) and 3-way (two features and the label) interactions.

In recent years, researchers have proposed many techniques to eliminate redundant features according to the above principle. Battiti proposed the Mutual Information based Feature Selection (MIFS) method [13], and then improved versions like MIFS-U [14] and mMIFS-U [15] were proposed. Hall proposed the Correlation-based Feature Selection (CFS) method [12], Ding and Peng proposed the minimum Redundancy-Maximum Relevance (mRMR) method [16], Bontempi and Meyer proposed a Causal filter selection method, called min-Interaction Max-Relevance (mIMR) [17], Fleuret used the Conditional Mutual Information Maximization (CMIM) to select feature [18]. These methods consider the problem of feature selection from different motivations and various solutions have been proposed. But the same point is that they are trying to find the optimal discriminative feature subset by considering to remove feature redundancy, where feature redundancy is computed by various pair-wise similarity measures, *i.e.* mutual information or conditional mutual information.

However, the used traditional pair-wise similarity metrics only consider the numerical values of given variables, but not the similarity of discriminative ability among them. Therefore, feature redundancy can not be measured correctly in terms of feature prediction ability by existing methods. How to measure feature (subset) redundancy is a fundamental problem, which should be reconsidered in the field of feature selection. In the application of classification, we consider a feature redundant only when its predictive power is redundant, not that its numerical value is similar with some selected features. Because two highly similar features are obviously not redundant to each other when the minor difference happen to be critical to the classification. In this paper, we proposed a novel method to measure feature (subset) redundancy by comparing features' predictive powers directly. Feature's prediction power is recorded by its instances' distribution explicitly, which includes clear-discerned instances and blur-discerned instances. Based on the predictive power distributions, a new feature redundancy metric is defined on the ground of comparing predictive powers. Furthermore, we proposed a novel algorithm named RESI (Redundant fEature Selection depending on Instance). Experiments on several benchmark microarray data sets demonstrate the outstanding performance of our proposed method.

Methods

Supervised redundant feature detection

Given a data set (X = [X_1, X_2, . . . , X_m], Y) with n instances represented by m features (genes), where X and Y are instances matrix and class label respectively. The task of classification is to tag instances with a label in low probability of error. Theoretically, having more features implies more discriminative power in classification. However, many features are relevant to each other and they have no contribution to classification, except to degrade performance of classifiers [19]. Thus, it is necessary to remove irrelevant and redundant features as more as possible, without losing information greatly. Generally, the task of feature selection in classification issue is formally defined as the process of selecting the optimal feature subset $S \subseteq X$ to have high classification performance.

Thus, qualities of relevant and non-redundant features of the selected subset are two key points considered in feature selection methods. In existing methods, predictive power of a feature subset has often been divided into computation of the interrelation score of individual features with the label. Due to computational limitation in high dimensional feature space, redundancy degree of a feature subset is usually estimated by pair-wise similarity measures, *i.e.* mutual information or conditional mutual information of two given variables.

However, to measure the feature subset redundancy is a fundamental problem, which should be reconsidered in the field of feature selection. Without including learning model (*i.e.* wrapper methods), measuring the redundancy degree of a feature subset as a whole is impractical. So, approximately measuring individual pair-wise feature similarity is a reasonable way which has been adopted by various existing methods. But, the previous pair-wise similarity measures only consider the numerical values of given features, but not the discriminative ability among them. Therefore, feature redundancy is not measured correctly. For instance, two highly similar features are usually considered as redundant to each other by the previous pair-wise similarity measures. But this is not correct when the minor difference of the two features happens to be critical to the classification.

In the context of classification, it is reasonable to address the feature set redundancy from the view of predictive power. So, we define the pair-wise feature redundancy as follow.

Theorem 1 In the context of classification, a feature $\mathrm{X_p}$ is redundant to a given feature $\mathrm{X_q}$ if its predictive power to label Y has already been expressed by feature $\mathrm{X_q}$.

As argued in Theorem 1, the similarity should be measured by comparing the distribution of predictive power between two features. Thus, the predictive power needs to be measured in a comparable way. More concretely, we need to record those instances which are clear-discerned by the given feature and which are not.

Theorem 2 Given a feature $\mathrm{X_p}$, an instance $\mathrm{X^u}$ is clear-discerned if the majority of its k nearest neighbour instances on feature $\mathrm{X_p}$ have the same class labels with that of $\mathrm{X^u}$.

The intuitive idea of Theorem 2 is inspired from the *k* Nearest Neighbor (*k*NN) classifier. We believe that Theorem 2 holds true no matter which classification model is applied finally. Based on the idea in Theorem 2, we define the neighborhood pUrity of instance X^u on a given feature X_p as below,

$$U(X_p^u|Y) = \frac{\sum_{X^v \in N_k(X_p^u)} \delta\,(Y^v == Y^u)}{k}, \tag{1}$$

where $\delta(\cdot)$ is the Kronecker delta function *i.e.* $\delta(\cdot) = 1$ when $Y^v == Y^u$ is true, otherwise $\delta(\cdot) = 0$, Y^u is the target label of instance X^u, X_p^u is the value of instance X^u on feature X_p, $N_k\ (X_p^u)$ is the neighborhood of X^u defined by the k closest instances X^v on feature X_p (X_p^v is close to X_p^u), k is a predefined parameters. Equation 1 measures how many instance labels are the same in the neighborhood of X^u via feature X_p.

The value of $U(X_p^u|Y)$ varies from 0 to 1. High $U(X_p^u|Y)$ means the corresponding instance has more neighbors with the same class. An instance is defined as clear-discerned when its $U(X_p^u|Y)$ is higher than a given threshold μ, *i.e.* $U(X_p^v|Y) > \mu$, otherwise it is blur-discerned.

By Equation 1, each feature's predictive power is recorded as a distribution of instances, *i.e.* the clear-discerned instances and the blur-discerned instances. Then, redundancy of two given feature can be measured by comparing the corresponding instance distributions directly. We define the REdundancy Measure depending on Instances (REMI) of feature X_q to feature X_p under label Y as follows.

$$REMI(X_p; X_q|Y) = \frac{\sum_{u=1}^{n} \delta(U(X_p^u|Y) \le \mu \ \&\ U(X_q^u|Y) \le \mu)}{\sum_{u=1}^{n} \delta(U(X_p^u|Y) \le \mu)} \tag{2}$$

The numerator of Equation 2 gives the count of instances which are blur-discerned on both feature X_p and feature X_q, the denominator is the blur-discerned instances which count on feature X_p. The value of $REMI(X_p;\ X_q|Y)$ varies from 0 to 1. It is 0, when all blur-discerned instances on feature X_p are clear-discerned on feature X_q, which means the predictive power of X_q is completely complementary to that of X_p and we consider their corresponding redundancy is zero. The value is 1 when no blur-discerned instance on feature X_p is clear-discerned on feature X_q, which means X_q has no predictive power contribution to X_p, i.e. X_q is redundant to X_p.

As to measure the predictive power of selected features, we use the metric of t-statistic. For binary classification, the definition of t-statistic on a feature p is given as:

$$t(p|Y) = \frac{\bar{p}^1 - \bar{p}^2}{var^1/n^1 + var^2/n^2} \tag{3}$$

where $\bar{p}^1$, n^1 and var^1 are the mean value of features, the number of examples and variance of one class, $\bar{p}^2$, n^2 and var^2 have the similar meaning for the other class.

The meaning of t-statistic value is intuitive, which measures the weighted distance between the centroid of one class and the other class on the feature p. It is usually believed that the value of $t(p|Y)$ largely represents the discriminative power of the giving feature, higher absolute score means greater discriminative power. A feature is usually regarded as irrelevant when its t-statistic value is trivial. The t-statistic measure has demonstrated as an effective feature selection method and widely used in the field of bioinformatics [20].

In total, we define the merit function of REMI as:

$$J_{REMI}(X_p|S, Y) = \mathrm{abs}(t(X_p|Y)) - \frac{1}{p-1}\sum_{i=1}^{p-1} REMI(X_p; X_i|Y) \tag{4}$$

which denotes the merit score of candidate feature X_p given the selected features S ($S = X_{1:p-1}$) and the label Y. Furthermore, we propose a novel algorithm named RESI (Redundant fEature Selection depending on Instance), which combines REMI with the sequential forward search strategy. The detail algorithm is described as in Algorithm 1.

Algorithm 1 The RESI Algorithm

```
Input: Feature set $X = [X_1, X_2, \ldots, X_m]$
       Target label $Y$
Output: Selected feature subset $S$

$F \Leftarrow X$;
$S \Leftarrow \emptyset$;
for all feature $X_i$ in $F$ do
    if abs($t(X_i|Y)$) $<\in$ then
        $F \Leftarrow F \setminus X_i$;
    end if
end for
$X_p \Leftarrow \arg max_{X_p \in F} \mathrm{abs}(t(X_p|Y))$;
$S \Leftarrow S \cup X_p$;
$F \Leftarrow F \setminus X_p$;
while pre-defined stopping criteria is not satified do
    $X_p \Leftarrow \arg max_{X_p \in F} J_{REMI}(X_p|S, Y)$;
    $S \Leftarrow S \cup X_p$;
    $F \Leftarrow F \setminus X_p$;
end while
```

The RESI algorithm works in two stages, including irrelevant features removing and redundant feature elimination. Firstly, the t-statistic value is computed for each feature. The irrelevant features, whose absolute t-statistic values are trival *i.e.* $\text{abs}(t(X_i|Y)) <\in$, are removed from the candidate set. Then, the most discriminative features with the highest $\text{abs}(t(X_p|Y))$ are selected. In the stage of redundant feature elimination, RESI starts with the current selected subset, and adds one important feature at a time. Given a selected feature subset of $p-1$ features S , a new feature X_p is chosen from the rest of the feature subset $X \setminus S$ by evaluating the merit function $J_{REMI}\ (X_p|S, Y)$. It will be terminated when pre-defined stopping criteria is satisfied. For example, the number of selected features is larger than a threshold or the score of $J_{REMI}\ (X_p|S, Y)$ is trivial.

In our experiments, the parameter $\in = 0.1$, the number of neighborhood instances $k = 3$, the threshold of clear-discerned instance $\mu = 0.66$.

Related works on redundant feature detection

As we mentioned in Section Introduction, the predictive power and the inner redundancy of the selected feature subset are two key points in feature selection methods. In recent years, researchers have proposed various feature selection methods [21-23], most of which are explained from the two aspects. Here, we give a short review on related feature selection methods using forward feature selection scheme. These methods are briefly described in Table 1.

Hall proposed the Correlation-based Feature Selection (CFS) method [12], where some correlation measures are used to evaluate the goodness of a subset by considering the individual predictive ability of each feature and the degree of correlation between them. The symmetrical uncertainty (a normalized version of mutual information), for discrete data, and the standard linear correlation, for continuous data, are used by Hall to measure $I(X_i;\ Y)$ and $I(X_i;\ X_j)$.

Ding and Peng proposed the minimum Redundancy-Maximum Relevance (mRMR) method in 2005 [16], which requires that selected discriminative features are maximally dissimilar to each other. mRMR is almost as the same as the MIFS method [13], except that the parameter β is set as $\frac{1}{p-1}$ in mRMR. Both MIFS and mRMR use mutual information to measure $I(X_i;\ Y)$ and $I(X_i;\ X_j)$. Kwak and Choi proposed an improvement to MIFS, called MIFS-U [14], which uses a re-weighted mutual information $\frac{I(X_i;Y)}{H(X_i)} I\left(X_p; X_i\right)$ to measure the feature redundancy. Without explicit claim, $I(\cdot;\cdot\ )$ is measured by the mutual information in the rest of paper.

Bontempi and Meyer proposed a causal filter selection method, called min-Interaction Max-Relevance (mIMR) [17]. Bontempi and Meyer try to maximize the mutual information between $X_{1:p}$ and Y directly. Due to the number of the subset of $X_{1:p}$ is $O(2^p)$, it is impractical to compute $I(X_{1:p};\ Y)$ in a precise way. Only low-order interactions are considered in the approximate solution given by mIMR. And in the final solution, the merit function of mIMR can also be divided into two parts: predictive power and feature set redundancy. Obviously, the difference between mRMR and mIMR is the two-way mutual information $I(X_i;\ X_p)$ is replaced by the three-way mutual information $I(X_i;\ X_p;\ Y)$, and $I(X_i;\ X_p;\ Y) = I(X_p;\ X_i) - I(X_p;\ X_i|Y)$.

Fleuret used the Conditional Mutual Information Maximization (CMIM) to select feature [18], which examines the information between a feature and the target, conditioned on each current feature. It is clear that CMIM is very similar with mIMR. The only difference is the sum function is replaced by the maximum function to measure the feature redundancy.

Yang and Moody proposed using Joint Mutual Information (JMI) to select feature [24], which tries to maximize joint mutual information $\sum_{i=1}^{p-1} I\left(X_p, X_i; Y\right)$. This is the information between the targets and a joint random variable, defined by pairing the candidate X_p with each current selected feature. But after deduction, the merit function of JMI is exactly equivalent to that of mIMR.

Vidal-Naquet and Ullman proposed another criterion used for computer vision, which is refereed as Informative

Table 1 Related feature selection methods

Methods	Predictive power	Redundancy measure	Merit function of candidate feature	Ref
CFS	$I(X_i;\ Y)$	$I(X_p;\ X_i)$	$\Delta\left(\frac{\sum_{i=1}^{p} I(X_i;Y)}{\sqrt{p+\sum_{i=1}^{p}\sum_{j=1}^{p} I(X_i;X_j)}}\right)$	[12]
MIFS	$I(X_p;\ Y)$	$I(X_p;\ X_i)$	$I(X_p;Y) - \beta \sum_{i=1}^{p-1} I(X_p;X_i)$	[13]
mRMR	$I(X_p;\ Y)$	$I(X_p;\ X_i)$	$I(X_p;Y) - \frac{1}{p-1} \sum_{i=1}^{p-1} I(X_p;X_i)$	[16]
MIFS-U	$I(X_p;\ Y)$	$\frac{I(X_i;Y)}{H(X_i)} I\left(X_p;X_i\right)$	$I(X_p;Y) - \beta \sum_{i=1}^{p-1} \left[\frac{I(X_i;Y)}{H(X_i)} I\left(X_p;X_i\right)\right]$	[14]
mIMR, JMI	$I(X_p;\ Y)$	$I(X_i;\ X_p;\ Y)$	$I(X_p;Y) - \frac{1}{p-1} \sum_{i=1}^{p-1} \left[I(X_i;X_p) - I(X_i;X_p\|Y)\right]$	[17],[24]
CMIM, IF	$I(X_p;\ Y)$	$I(X_i;\ X_p;\ Y)$	$I(X_p;\ Y) - \max_{i\ \in[1:p-1]}[I(X_p;\ X_i) - I(X_p;\ X_i\|Y)]$	[18],[25]

Fragments (IF) [25]. The authors motivate the criterion $\min_{i\in[1:p-1]}[I(X_p, X_i; Y) - I(X_i; Y)]$ by noting that it measures the predictive ability gain of combining a new feature X_p with each existing feature X_i, over simply using X_i by itself. The X_i with the least "gain" from being paired with X_p is taken as the score for X_p. Interestingly, using the chain rule $I(X_p, X_i; Y) = I(X_i; Y) + I(X_p; Y|X_i)$, therefore IF is equivalent to CMIM.

From the short review, we can easily find the common point of these methods. Although the motivations are various, the merit functions of all these methods are divided into two parts: predictive power and feature redundancy. And due to the practical limitation, one certain pair-wise similarity measure is adopted to compute the predictive power and feature redundancy. However, the traditional pair-wise similarity measures, *i.e.* mutual information or conditional mutual information, only consider the numerical values of given variables, but not the the similarity of discriminative ability between them. Therefore, feature redundancy can not be measured correctly in existing methods.

Data sets

There are fourteen data sets used in our study, which are listed in Table 2. The data set of Arcene-NIPS2003 is gathered from the NIPS'03 feature selection competition [26], Breast-Duke is reported by West [27], all other data sets are downloaded from the Kent Ridge Bio-medical Dataset [28]. All these data sets have relative big feature/instance ratio, and the feature numbers are no less than 2,000. For the missing values in some existing data sets, they are replaced by the corresponding means. For the data set of OvarianQStar, only the first 373,401 features are used.

Experimental settings

We use the stratified 10-fold cross-validation procedure, where each data set is split into ten subsets of equal size. Each subset is used as test set once, and the corresponding left subsets are combined together and used as training set. Within each cross-validation fold, the gene expression data are standardized. The expressions of the training set are transformed to zero mean and unit standard deviation across samples, and the test set are transformed according to the means and standard deviations of the corresponding training set. The Irani's MDL method is applied when discretization is required [29]. The 10-fold cross-validation is repeated 10 times, which is also denoted as the 10 × 10 cross-validation measuring procedure.

Table 2 Experimental data sets

Data sets	Instances	Class ratio	Features
Arcene-NIPS2003	900	398/502	10,000
Breast	97	46/51	24,481
Breast-Duke	44	21/23	7,129
CNS	60	21/39	7,129
Colon	62	22/40	2,000
DLBCL-Stanford	47	23/24	4,026
DLBCL-Tumor	77	19/58	6,817
DLBCL-NIH	240	102/138	7,399
Leukemia	72	25/47	7,129
Lung	181	31/150	12,533
Lung-Michigan	96	10/86	7,129
Lung-Ontario	39	15/24	2,880
OvarianPBSII	253	91/162	15,154
OvarianQStar	216	95/121	373,401

We should note that the 10 × 10 cross-validation measuring procedure is more reliable than the randomized re-sampling testing strategy and the leave-one-out cross-validation due to the correlations between the test and training sets, some detailed discussions can be found at [30] Even in the small sample problem like gene expression data, 10 × 10 cross-validation is still one of the most reliable measuring way [31].

The final classification performance is recorded by the Balanced ACCuracy (BACC), which is defined as follows.

$$\begin{aligned} BAAC &= \frac{1}{2}\left(sensitivity + specificity\right) \\ &= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \end{aligned} \tag{5}$$

where TP, TN, FP, and FN, stand for the number of true positive, true negative, false positive, and false negative samples, respectively. Without explicit clarification, all the scores are averaged on 10 × 10 cross-validation.

To make conclusions sound, six widely used classifiers are used, including Support Vector Machine (SVM) with linear kernel and $c = 1$, non-linear support vector machine using Sequential Minimal Optimization (SMO) and polynomial kernel, k Nearest Neighbor (kNN) with $k = 3$, Logistic Regression (LR), Naïve Bayes classifier (NB) and decision tree with J48 algorithm (J48). All these classifiers are trained on the training set to predict the label of the test samples on the same cross-validation partition.

The algorithms are implemented in JAVA language based on WEKA [32], and carried out on a DELL PC workstation with 24 × X5680 3.33GHz CPU and 64G RAM.

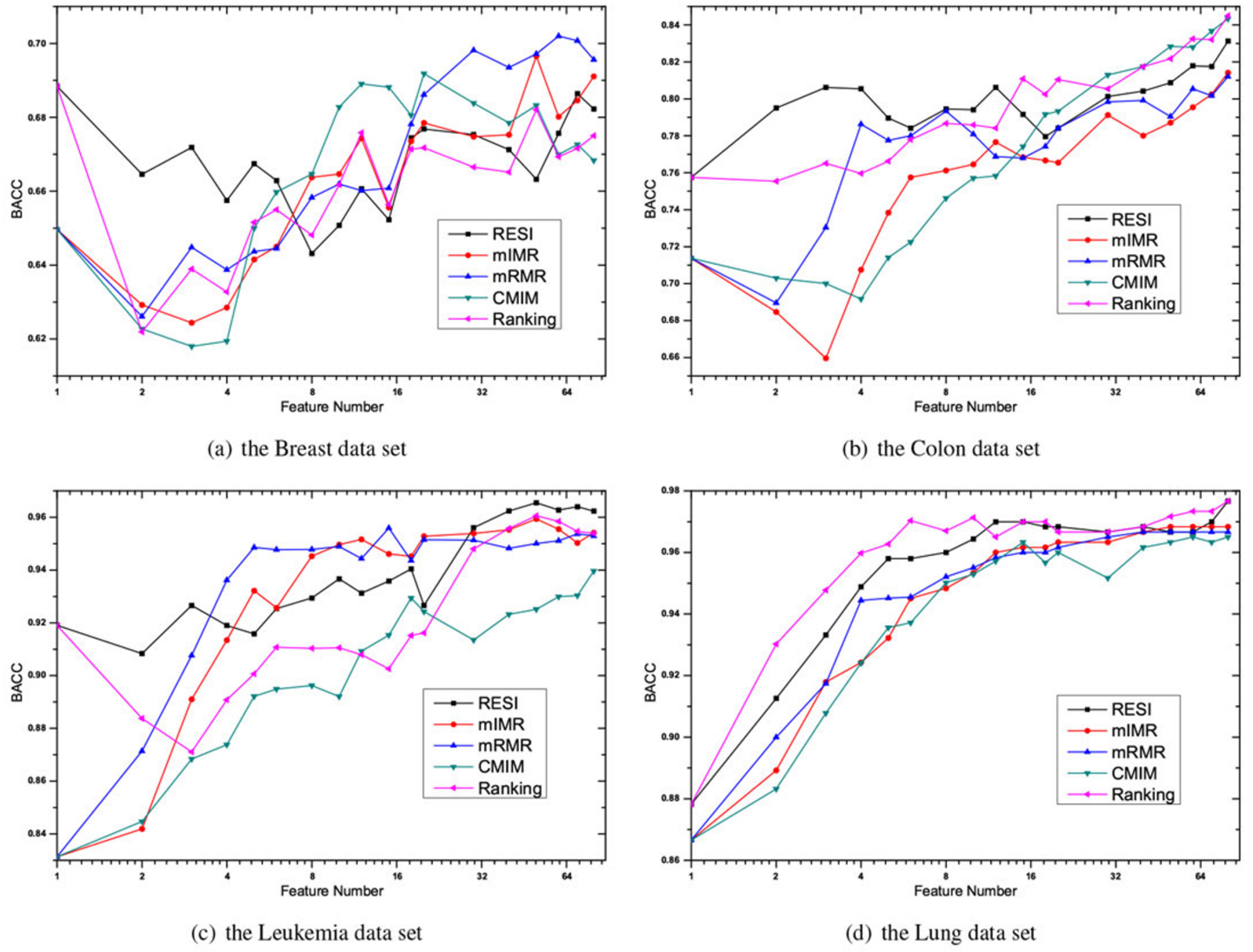
Results and discussion

In order to examine the performance of our proposed method, three state-of-the-art feature selection methods, mIMR, mRMR and CMIM, are used to compare with RESI. Additional, a feature Ranking method using absolute t-statistic score is also used as baseline. The parameter of selected feature number has great influence on the performance comparison. We vary the selected dimension from 1 to 80, which are 1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 18, 20, 30, 40, 50, 60, 70 and 80 in detail. Higher dimension is not included because the number of relevant genes (whose $I(X_i; Y) > 0$) is limited on some data sets. On each data set and each selected feature dimension, six widely used classifiers *i.e.* SVM, SMO, LG, *k*NN, NB and J48 are applied to examine the performances with the procedure of 10 × 10 cross-validation. The detailed comparative BACC results with *k*NN classifier are plotted in Figure 1. Due to the length limitation of the paper, only results on four representative data sets, including the data set of Breast, Colon, Leukemia and Lung, are included in Figure 1, which are widely used by previous researchers [16,17]. The comparative BACC results averaged on all fourteen data sets are plotted in Figure 2. Note, the abscissas of Figure 1 and Figure 2 use the Log coordinate and the Log base is 2.

From Figure 1 and Figure 2, it's clear there is no method has overwhelming performance over all classifiers, although the plotted lines in Figure 2 fluctuate much more gently than the drawings on individual data set. However, our RESI is relatively the best one among all the feature selection methods. This is especially obvious by using the classifier of LR, and we will demonstrate the point by t-test later.

The performances of mIMR, mRMR and CMIM are often tied, and mRMR is the best one in most cases. We suggest it's because all these methods use mutual information to represent the predictive power, and the traditional mutual information of mRMR is the most robust one to estimate feature redundancy in our experiments.

It's also interesting to find that the performances of RESI and Ranking are obviously better than that of others, when the selected dimension is small, *i.e.* below 5. This is true with almost all classifiers. We believe the superiority is caused by the difference of two metrics, mutual information and the absolute t-statistic, used to represent the predictive power. The absolute t-statistic measures the weighted distance between the centroid of two classes, which is directly related to the discriminative power of the



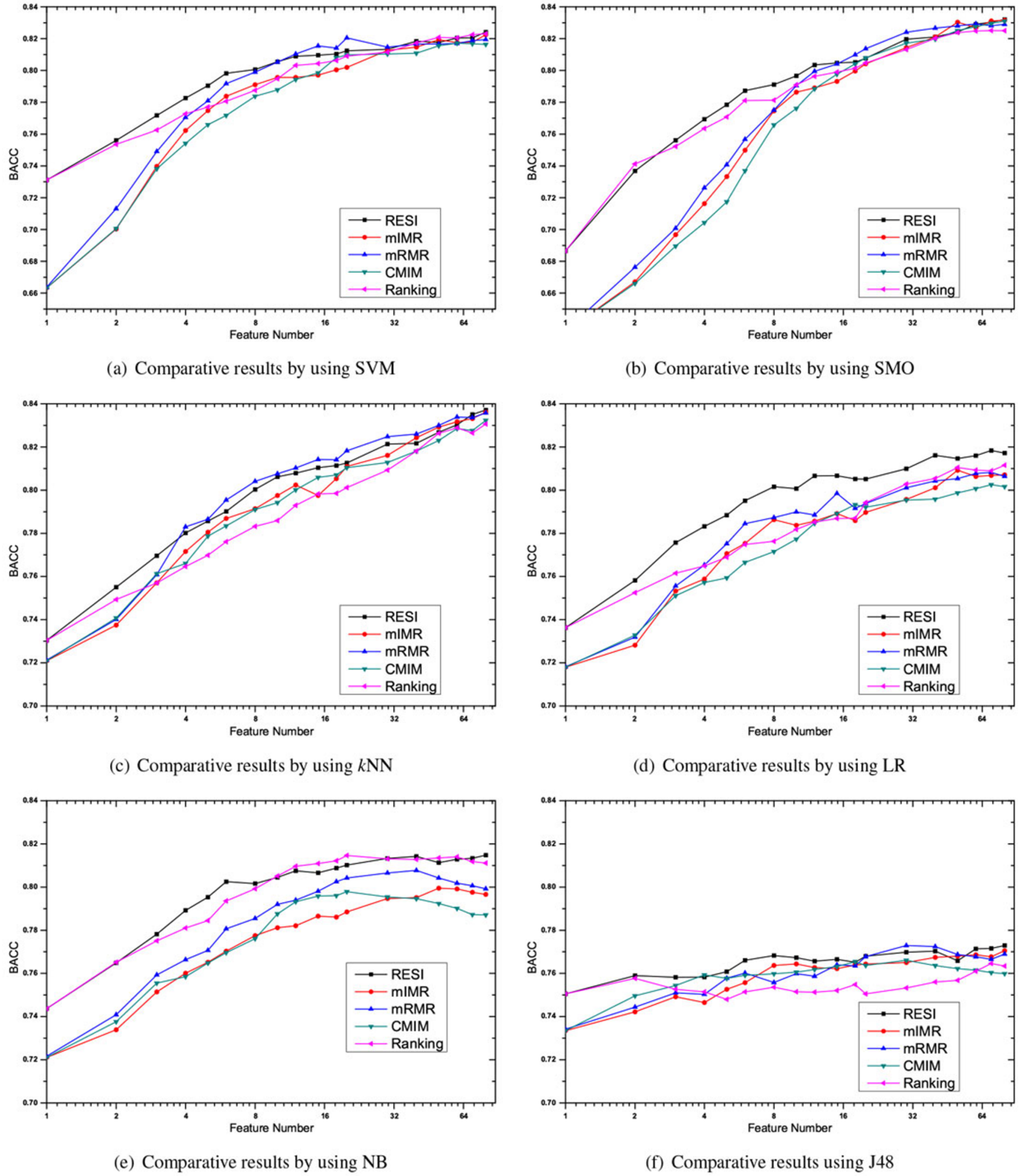
(a) the Breast data set

(b) the Colon data set

(c) the Leukemia data set

(d) the Lung data set

Figure 1 Comparative BACC results on four selected data sets by using *k*NN



(a) Comparative results by using SVM

(b) Comparative results by using SMO

(c) Comparative results by using *k*NN

(d) Comparative results by using LR

(e) Comparative results by using NB

(f) Comparative results using J48

Figure 2 Averaged comparative BACC results on fourteen data sets by using different classifiers

giving feature. On the contrary, mutual information represents the general information between two variables. So from our experimental results, the absolute t-statistic performs better than mutual information to represent the predictive power.

Investigating the difference between RESI and Ranking, we find that their performances are similar when the selected dimension is small. This is because the absolute t-statistic is used to represent the predictive power by both of them. And when only a few features

Table 3 Comparative t-test results (wins/ties/losses) summarized on data sets and classifiers

	mIMR	mRMR	CMIM	Ranking	Total
RESI	40-33-11	36-37-11	41-33-10	31-36-17	148-139-49
mIMR	-	9-42-33	31-28-25	20-30-34	71-133-132
mRMR		-	36-36-12	23-32-29	103-147-86
CMIM			-	18-27-39	65-124-147
Ranking				-	119-125-92

are selected, the influence of feature redundancy is not obvious. But as the selected dimension grows, the impact of feature redundancy becomes more and more critical to the BACC performance. Eliminating redundant features is meaningful when the feature dimension is not small, which is coincided with Figure 2, where RESI is much better than Ranking when the dimension grows.

Paired two-side t-test is also used to examine the differences between those methods. The corresponding t-test results are showed in Table 3. Each cell (W/T/L) in Table 3 summarizes over all data sets and classifiers the wins/ties/losses in BACC (at the significance level of 0.05) comparing various feature selection methods each other. The last column of Table 3 gives the overall W/T/L values summarized on corresponding compared methods. From Table 3, it's clear that RESI is the best feature selection method.

Conclusions

Redundant feature selection is an important topic in the field of bioinformatics. Here, we propose a novel redundant feature subset measure REMI by comparing feature predictive powers directly, which is recorded by its instance distribution explicitly including clear-discerned instances and blur-discerned instances. Furthermore, a novel feature selection method RESI based on REMI was proposed. Experimental results on benchmark microarray data sets demonstrate that RESI performs better than the state-of-the-art algorithms like mRMR on fourteen benchmark data sets.

Future works include improving its efficiency and applying it to more scientific fields.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors read and approved the final manuscript. GZL proposed the algorithm and wrote this paper. XQZ jointly proposed the paper idea, collected the data, implemented the computation on the datasets and checked the manuscript.

Acknowledgements

This work was supported by the Natural Science Foundation of China under grant nos. 61105053, 61273305 and 61463033, the Nature Science Foundation of Jiangxi Province under grant no. 20132BAB201043, and China Postdoctoral Science Foundation under grant no. 2013M541540.

Declarations

Funding for open access charge: Natural Science Foundation of China. This article has been published as part of *BMC Medical Genomics* Volume 7 Supplement 2, 2014: IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013): Bioinformatics in Medical Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedgenomics/supplements/7/S2.

Authors' details

[1]Computer Center, Nanchang University, 999 Xuefu Road, Nanchang, 330029, China. [2]Department of Control Science & Engineering, Tongji University, 4800 Cao An Road, Shanghai, 201804, China. [3]The Key Laboratory of Embedded System & Service Computing, Tongji University, 4800 Cao An Road, Shanghai, 201804, China.

Published: 22 October 2014

References

1. Ge L, Du N, Zhang A: **Finding informative genes from multiple microarray experiments: A graph-based consensus maximization model.** *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on* 2011, 506-511.
2. Dougherty ER: **Small sample issue for microarray-based classification.** *Comparative and Functional Genomics* 2001, **2**:28-34.
3. Zheng S, Liu W: **Selecting informative genes by lasso and dantzig selector for linear classifiers.** *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on* 2010, 677-680.
4. Zeng XQ, Li GZ: **Incremental partial least squares analysis of big streaming data.** *Pattern Recognition* 2014, **47(11)**:3726-3735.
5. You M, Li GZ: **Feature selection for multi-class problems by using pairwise-class and all-class techniques.** *International Journal of General Systems* 2011, **40(4)**:381-394.
6. Li GZ, Meng HH, Yang MQ, Yang JY: **Combining support vector regression with feature selection for multivariate calibration.** *Neural Computing and Applications* 2009, **18(7)**:813-820.
7. Shao H, Li G, Liu G, Wang Y: **Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine.** *Science China Information Sciences* 2013, **56(5)**:1-13.
8. Li GZ, Sun S, You M, Wang YL, Liu GP: **Inquiry diagnosis of coronary heart disease in chinese medicine based on symptom-syndrome interactions.** *Chinese Medicine* 2012, **7(1)**:9.
9. You M, Liu J, Li GZ, Chen Y: **Embedded feature selection for multi-label classification of music emotions.** *International Journal of Computational Intelligence Systems* 2012, **5(4)**:668-678.
10. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *The Journal of Machine Learning Research* 2003, **3**:1157-1182.
11. Zeng XQ, Li GZ, Yang JY, Yang MQ, Wu GF: **Dimension reduction with redundant gene elimination for tumor classification.** *BMC Bioinformatics* 2008, **9(Suppl 6)**:8.
12. Hall MA: **Correlation-based feature selection for discrete and numeric class machine learning.** *Proceedings of the 2000 International Conference on Machine Learning (ICML'00)* 2000, 359-366.
13. Battiti R: **Using mutual information for selecting features in supervised neural net learning.** *IEEE Transactions on Neural Networks* 1994, **5(4)**:537-550.
14. Kwak N, Choi CH: **Input feature selection for classification problems.** *IEEE Transactions on Neural Networks* 2002, **13(1)**:143-159.
15. Novovičová J, Somol P, Haindl M, Pudil P: **Conditional mutual information based feature selection for classification task.** In *Progress in Pattern Recognition, Image Analysis and Applications. Volume 4756.* Springer; 2007:417-426.
16. Peng H, Long F, Ding C: **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005, **27(8)**:1226-1238.
17. Bontempi G, Meyer P: **Causal filter selection in microarray data.** *Proceedings of the 2010 International Conference on Machine Learning (ICML'10)* 2010, 95-102.
18. Fleuret F: **Fast binary feature selection with conditional mutual information.** *The Journal of Machine Learning Research* 2004, **5**:1531-1555.
19. Zeng XQ, Li GZ: **Dimension reduction for p53 protein recognition by using incremental partial least squares.** *IEEE Transactions on NanoBioscience* 2014, **13(2)**:73-79.

20. Liu H, Li J, Wong L: **A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns.** *Genome informatics* 2002, **13**:51-60.
21. Brown G: **A new perspective for information theoretic feature selection.** *Proceedings of the 2009 International Conference on Artificial Intelligence and Statistics (ICAIS'09)* 2009, **5**:49-56.
22. Liu H, Liu L, Zhang H: **Feature selection using mutual information: An experimental study.** *Proceedings of the 2008 Pacific Rim International Conference on Artificial Intelligence (PRICAI'08)* 2008, 235-246.
23. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A: **A review of feature selection methods on synthetic data.** *Knowledge and Information Systems* 2013, **34(3)**:483-519.
24. Yang HH, Moody J: **Data visualization and feature selection: New algorithms for nongaussian data.** *Advances in Neural Information Processing Systems* 1999, **12**:687-693.
25. Vidal-Naquet M, Ullman S: **Object recognition with informative features and linear classification.** *Proceedings of the 2003 IEEE International Conference on Computer Vision (ICCV'03)* 2003, 281-288.
26. Guyon I, Li J, Mader T, Pletscher PA, Schneider G, Uhr M: **Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark.** *Pattern Recognition Letters* 2007, **28(12)**:1438-1444.
27. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proceedings of the 2001 National Academy of Sciences* 2001, 11462-11467.
28. Li J, Liu H, Wong L: **Mean-entropy discretized features are effective for classifying high-dimensional biomedical data.** *Proceedings of the 3rd ACM SIGKDD Workshop on Data Mining* 2003, 17-24.
29. Fayyad U, Irani K: **Multi-interval discretization of continuous-valued attributes for classification learning.** *Proceedings of the 1993 International Joint Conference on Articial Intelligence (IJCAI'93)* 1993, 1022-1027.
30. Dietterich TG: **Approximate statistical tests for comparing supervised classification learning algorithms.** *Neural Computation* 1998, **10**:1895-1923.
31. Molinaro AM, Simon R, Pfeiffer RM: **Prediction error estimation: a comparison of resampling methods.** *Bioinformatics* 2005, **21(15)**:3301-3307.
32. Witten IH, Frank E: **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann Publishers Inc., San Francisco;, 2.

doi:10.1186/1755-8794-7-S2-S5
Cite this article as: Zeng and Li: **Supervised redundant feature detection for tumor classification.** *BMC Medical Genomics* 2014 **7**(Suppl 2):S5.