BioData Mining

METHODOLOGY Open Access

A feature selection method based on multiple kernel learning with expression profiles of different types

Wei Du[1], Zhongbo Cao[1,3], Tianci Song[1], Ying Li[1*] and Yanchun Liang[1,2*]

* Correspondence: liying@jlu.edu.cn; ycliang@jlu.edu.cn
[1]College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun, 130012, China
Full list of author information is available at the end of the article

Abstract

Background: With the development of high-throughput technology, the researchers can acquire large number of expression data with different types from several public databases. Because most of these data have small number of samples and hundreds or thousands features, how to extract informative features from expression data effectively and robustly using feature selection technique is challenging and crucial. So far, a mass of many feature selection approaches have been proposed and applied to analyse expression data of different types. However, most of these methods only are limited to measure the performances on one single type of expression data by accuracy or error rate of classification.

Results: In this article, we propose a hybrid feature selection method based on Multiple Kernel Learning (MKL) and evaluate the performance on expression datasets of different types. Firstly, the relevance between features and classifying samples is measured by using the optimizing function of MKL. In this step, an iterative gradient descent process is used to perform the optimization both on the parameters of Support Vector Machine (SVM) and kernel confidence. Then, a set of relevant features is selected by sorting the optimizing function of each feature. Furthermore, we apply an embedded scheme of forward selection to detect the compact feature subsets from the relevant feature set.

Conclusions: We not only compare the classification accuracy with other methods, but also compare the stability, similarity and consistency of different algorithms. The proposed method has a satisfactory capability of feature selection for analysing expression datasets of different types using different performance measurements.

Background

With the development of transcriptomics research, especially the widely used high-throughput microarray chip and RNA sequencing technology, a large number of transcriptome data have been obtained by measuring the expressions of genes or miRNAs simultaneously. Researchers can acquire these different expression data from several public databases, such as Gene Expression Omnibus (GEO) [1], Stanford Microarray Database (SMD) [2], ArrayExpress [3] and The Cancer Genome Atlas (TCGA) [4]. TCGA is the largest cancer genome sequencing project, which plan to sequence and organize 10,000 cancer genomes, along with other matching omics data types, covering 25 cancer types [5]. Developing effective and robust methods to extract the subset of

BioMed Central

informative features from expression data remains a challenge and crucial problem. Feature selection technology has been studied and applied proverbially in pattern recognition, statistics analysis, data mining and machine learning [6]. In the last decade, feature selection technology has become an important tool for expression data analysis in the field of bioinformatics, such as cancer classification, biological network inference, expression correlation analysis and disease biomarker identification [7]. The features (mRNAs or miRNAs) of given expression data can be broadly categorized into three major types: relevant features, redundant features and irrelevant features [8].

In general, most feature selection methods can be divided into three categories: filter methods, wrapper methods, and embedded methods [7]. These categories depend on the combination modality of feature selection search and the construction of the classification model. **Filtering methods**, which are independent of the classifier, select relevant features only dependent the intrinsic properties of expression data. Glaab et al. applied an ensemble filter method which combines several selection schemes to an ensemble feature ranking [9]. Cai et al. proposed a feature weighting algorithm to estimate the feature weights through local approximation rather than global measurement. Experimental results on both synthetic and real microarray datasets validated that the algorithm was effective, when combining the proposed method with classic classifiers [10]. Cao et al. proposed a filtering feature selection method for paired microarray expression data analysis [11].

In **wrapper approaches**, the classification scores for features by a classifier are measured in the selection process and the step of feature selection depends on the classifier. So far, many wrapper feature selection methods have been proposed and used for expression data analysis. Mukhopadhyay et al. combined a multi-objective genetic algorithm and SVM classifier as a wrapper for evaluating the chromosomes that encode miRNA feature subsets [12]. Maulik et al. presented a fuzzy preference based rough set method for feature selection from gene expression data of microarray. Compared with signal-to-noise ratio and consistency based Feature Selection methods, experimental results showed that the method was effective in extracting gene markers [13].

In **embedded approaches**, the step of selecting an optimal feature subset is built into the classifier construction and the selecting can be seen the process combined space of feature subsets and hypotheses. With the increase of available expression data sources, several embedded feature selection methods have been presented to analyze expression data. Chen et al. proposed a feature selection approach using the information provided by the separating hyperplane and support vectors [14]. Mao et al. proposed a unified feature selection framework based on a generalized sparse regularizer for measuring the performance of multivariate [15]. Li et al. proposed a new feature selection algorithm called feature weighting as regularized energy-based learning. The experiments using microarray data demonstrated that the ensemble method, when using the L2 regularizer outperforms other algorithms in stability while providing comparable classification accuracy [16]. Kursa compared four state-of-the-art Random Forest-based feature selection methods in the gene selection context on microarray datasets, and found when the number of consistently selected genes was considered, the Boruta algorithm was the best one [17]. Yousef et al. developed a method for selecting significant genes, which uses K-means to identify correlated gene clusters and applies the scores of those gene clusters for the purpose of classification [18]. Tang et al. presented a two-stage

Recursive Feature Extraction (RFE) algorithm, which can effectively eliminate most of the irrelevant, redundant and noisy genes, and select informative genes in different stages [8]. Niijima et al. suggested a recursive feature elimination model based on Laplacian linear discriminant analysis for feature selection [19]. However, these methods based on RFE may obtain satisfactory performance on hundreds of features. Such a large number of features (mRNAs or miRNAs) are difficult to apply to several fields, such as clinical diagnosis of cancer or experiments of identifying cancer biomarkers.

In recent years, several **hybrid feature selection approaches** have been also proposed for expression data analysis. Chuang et al. proposed a feature selection method, which combines an improved particle swarm optimization with the K-nearest neighbor method and support vector machine classifiers [20]. Mundra et al. developed a hybrid feature selection method by combining the filter method of minimum-redundancy maximum-relevancy (MRMR) and the wrapper method of support vector machine recursive feature elimination (SVM-RFE) [21]. Du et al. proposed a multi-stage feature selection method for microarray expression data analysis [22].

Though most of above methods can eliminate the irrelevant genes and rank informative genes effectively, they are only suitable for expression data from one type of expression profile. Most of the above methods construct the feature selection model based on one type of expression data directly, but they rarely consider the effectiveness and stability on expression data from different types of transcriptome. In this paper, we propose a novel two-stage feature selection method which uses multiple kernel learning (MKL) [23, 24] combines a forward feature selection procedure to select the relevant feature subset, eliminate redundant features and select compact feature subsets. We simplify our proposed method as Simple MKL-Feature Selection (SMKL-FS), which eliminates irrelevant features and selects relevant features by the score of individual feature, and eliminates redundant features by the forward selection procedure in two stages.

One objective of feature selection is to avoid overfitting and improve the performance of classifier [7]. Overfitting is one of challenging problems on gene expression data which have characteristic of high dimensional and small sample. So, we used following processing to decrease the influence of overfitting on small samples. Firstly, we use the SimpleMKL method, which solves the MKL problem through a primal formulation involving a weighted l2-norm regularization. The regularization part adds a cost term for bringing in more features with the objective function. Hence, regularization can shrink the coefficients of many variables to zero and decrease the overfitting. Secondly, we used a sequential forward selection (SFS) method which belonged to deterministic methods and have lower overfitting risk than randomized methods [7]. In addition, we used cross validation in performance measurement part to identify these methods, which may have poor performance caused by overfitting training on several datasets.

In the following part, we outline the main steps of SMKL-FS. Firstly, we measure the relevance between features and classify samples by using the optimizing function of MKL. More specifically, we use an iterative gradient descent process to perform the optimization both on the parameters of SVM and kernel confidence, and obtain the optimizing function of each feature. Then, we select the relevant features set by sorting the optimizing function of each feature. Furthermore, we apply an embedded scheme of forward selection to detect the compact feature subsets from the relevant features set. Different from wrapper approaches, which convolve with a classifier and minimize the classification error of the dependent classifiers, we use optimizing function of MKL instead of classification error to carry out the embedded process. The idea of this process is similar as the minimum-redundancy process in mRMR [25]. Except for evaluating the classification accuracy of the method, we measure the performances of different feature selection algorithms through measuring the stability of feature space on different samples in the same type of data, the similarity with other methods and consistency between expression data of miRNA and mRNA.

The main characteristics of our proposed algorithm include: (i) a novel feature selection method for identifying gene signatures based on multiple kernel learning focusing on multiple types of expression data, such as mRNA microarray, mRNA sequencing and miRNA sequencing; (ii) an evaluattion performance of different methods by using classification accuracy, stability of feature space, similarity with other methods and consistency between expression data of miRNA and mRNA. Experimental results show that the proposed method has a satisfactory capability of feature selection for different expression datasets analysis compared to other state of art feature selection approaches.

Results

For measuring the performance of embedded method, we use three kernel functions, linear kernel $K(x_i, x) = (x_i, x)$, radial basis function kernel $K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2}\right)$ and polynomial kernel $K(x_i, x) = [(x_i, x) + 1]^2$. In a practical application, different kernels can combined. The features are selected and evaluated using 10-fold Cross-Validation (CV) on a variety of datasets through different feature selection methods including SVM-RFE [26], SVM-RCE [18], mRMR [25], IMRelief [10], SlimPLS [27] and SMKL-FS. We measure the performances of different feature selection algorithms through evaluating the classification accuracy of feature combination, also measuring the stability of feature space on different samples in the same type of data and the similarity with other methods.

Data sources and pre-processing

In this paper, three types of expression data are used to measure the performance of feature selection methods. We only use the paired samples in expression datasets which include tumor and adjacent non-tumor tissues. The datasets of mRNA microarray are obtained from Gene Expression Omnibus (GEO) [1], the datasets of mRNA sequencing and miRNA sequencing are downloaded from The Cancer Genome Atlas (TCGA) [4]. Eight types of cancer on microarray datasets are used in this article, and each type of cancer contains several datasets (series in GEO). Table 1 gives the more detailed information of the eight cancer types of mRNA microarray datasets from GEO and Table 2 shows the more detailed information of the eight cancer types from TCGA.

For using these expression data to measure the performance of different feature selection methods, the downloaded and reorganized data from GEO and TCGA have been converted in our defined data format and preprocessed through the following processes. Firstly, the missing values of each expression dataset are estimated. If the missing values of one mRNA (or miRNA) are less than 20% of all samples, these missing values are estimated using the local least squares imputation (LLSimpute) method [28]. Then, the different probes of the same mRNA (or miRNA) are merged by the maximum expression value of these probes for each sample. After these processes, these datasets

Table 1 The detailed information of mRNA microarray datasets

Cancer Type	Datasets ID	Number of Samples
Liver	GSE5364, GSE22058, GSE14520, GSE12941	132
Pancreatic	GSE15471, GSE16515, GSE22780	63
Lung	GSE5364, GSE19804, GSE22058, GSE10072, GSE7670, GSE2514	249
Colon	GSE5364, GSE8671, GSE25070, GSE21510, GSE23878, GSE18105	70
Gastric	GSE13911, GSE13195, GSE5081, GSE19826	93
Breast	GSE5364, GSE15852, GSE10810, GSE16873, GSE5764, GSE14548	113
Thyroid	GSE5364, GSE3678	23
Prostate	GSE6919, GSE6956, GSE17951	88

are normalized by median absolute deviation (MAD) method to make all the samples have similar background [29]. The normalization of different microarrays is applied by housekeeping gene as performed in previous article [30].

Performance measurement of feature space

The performance measurement of feature space is important for evaluating different feature selection algorithms. Most of the state of art algorithms only validate their performance through the classification accuracy [26] or classification error [31] on selected feature set by a classifier *C*. The classification accuracy and classification error are defined as follows respectively:

$$\begin{aligned}&\text{Accuracy} = \frac{TP + TN}{FN + TP + TN + FP}\\&\text{Classification Error} = \frac{FN + FP}{FN + TP + TN + FP}\end{aligned} \tag{1}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives. However, only computing the classified ability of selected features could not reflect the performance of feature selection algorithms roundly.

In this paper, we measure the performances of different feature selection algorithms through evaluating the classification accuracy of single features and features combination, also measuring the stability of feature space on different samples in the same type of data, the similarity with other methods and consistency between expression

Table 2 The detailed information of mRNA Sequencing and miRNA Sequencing datasets

Cancer Type	Number of Samples
KIDNEY[1]	88
BRCA	71
LUNG[2]	47
HNSC	37
LIHC	46
PRAD	43
STAD	29
THCA	56

1: KIDNEY contains KIRC and KIRP
2: LUNG contains LUSC and LUAD

data of miRNA and mRNA. We select and evaluated features using 10-fold Cross-Validation (CV) on these datasets mentioned above through different feature selection methods, SVM-RFE [26], SVM-RCE [18], mRMR [25], IMRelief [10], SlimPLS [27], OSFS [32], FGM [33] and our method SMKL-FS. Firstly, for each testing dataset, we randomly selected 90% as training dataset and other 10% as test dataset. Repeating the selection process 10 times, we can obtain a collection of 10 groups contained training and test samples. In order to ensure fairness, we select feature subset using each feature selection method on training samples of the same 10 groups. Then, for the ten selected features from different methods, we evaluate them according to the above criterions.

Classification accuracy of features combination

For two feature sets S_n^1 and S_n^2, and the above classifier C, we consider the feature space of S_n^1 is more *effective*, if the classification accuracy on feature set S_n^1 is higher than that on S_n^2 by using classifier C. If the method M^1 generates a series of feature subsets in S_n^1 : $S_1^1 \subset S_2^1 \subset \ldots S_{n-1}^1 \subset S_n^1$ and the method M^2 generates a series of feature subsets in S_n^2 : $S_1^2 \subset S_2^2 \subset \ldots S_{n-1}^2 \subset S_n^2$. For each $k(1 \leq k \leq n)$, we compute the classification accuracy on S_k^1 and S_k^2 as same as [8]. If the average of these classification accuracies on S_n^1 is higher than that on S_n^2, we consider the method M^1 is better than M^2 in ***mean effectiveness***. If the maximum of these classification accuracies on S_n^1 is higher than that on S_n^2, we consider the method M^1 is better than M^2 in ***max effectiveness***.

In our verification, we set the n of feature set S_n^1 as 10, and compare the ***effectiveness*** of feature spaces from different methods using SVM classifier. For the feature subsets in S_{10}^1 : $S_1^1 \subset S_2^1 \subset \ldots S_9^1 \subset S_{10}^1$ generated by method M^1, we compute the classification accuracy on S_k^1 for every $k(1 \leq k \leq 10)$. Then the ***mean effectiveness*** and ***max effectiveness*** of method M^1 are measured by the average and maximum classification accuracies on S_{10}^1. The results of ***mean effectiveness*** and ***max effectiveness*** on three types of datasets through different methods are shown in Tables 3, 4 & 5 and Additional file 1: Table S1, respectively.

The ***mean effectiveness*** and ***max effectiveness*** of SMKL-FS are better than those from other methods for most datasets of miRNA sequencing, mRNA microarray data and little less than mRMR on mRNA sequencing data. The good performance of mRMR [25] on gene expression data may attribute to the method designed specifically

Table 3 The results of mean effectiveness on mRNA microarray (top 10)

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	OSFS	FGM	SMKL-FS
Liver	0.913	0.860	**0.965**	0.825	0.831	0.750	0.867	0.963
Pancreatic	0.689	0.777	*0.818*	0.784	0.673	0.707	0.729	0.804
Lung	0.731	0.786	0.942	0.814	0.708	0.704	0.860	*0.964*
Gastric	0.614	0.724	0.688	0.566	0.636	0.533	0.640	*0.760*
Colon	0.736	0.888	0.941	0.803	0.794	0.682	0.812	*0.951*
Breast	0.745	0.776	0.832	0.545	0.693	0.728	0.769	*0.854*
Thyroid	0.835	0.897	0.838	0.633	0.743	0.517	0.802	*0.922*
Prostate	0.577	*0.762*	0.750	0.560	0.682	0.629	0.679	0.717
Mean	0.730	0.809	0.847	0.691	0.720	0.656	0.770	*0.867*

Table 4 The results of mean effectiveness on mRNA Sequencing (top 10)

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	OSFS	FGM	SMKL-FS
KIDNEY	0.912	0.952	**0.965**	0.949	0.898	0.914	0.951	0.957
BRCA	0.938	0.982	0.973	0.953	0.871	0.934	0.928	*0.984*
LUNG	0.957	0.977	0.993	0.932	0.942	0.867	0.931	**0.997**
HNSC	0.930	0.949	*0.983*	0.908	0.844	0.900	0.977	0.948
LIHC	0.893	0.937	**0.962**	0.919	0.900	0.798	0.952	0.958
PRAD	0.932	0.928	**0.971**	0.893	0.779	0.764	0.966	0.953
STAD	0.907	0.895	**0.970**	0.945	0.758	0.848	0.898	0.963
THCA	0.945	0.954	**0.975**	0.933	0.883	0.844	0.903	0.970
Mean	0.927	0.947	**0.974**	0.929	0.859	0.859	0.938	0.966

for this type of data. We also see that FGM [33] is the best common method, which has satisfactory performance on different type of gene expression data. The results of accuracy of each $S_1^1, S_2^1, \ldots, S_9^1, S_{10}^1$ on three types of datasets for different methods are shown (See Additional file 2: Figure S1, Additional file 3: Figure S2 and Additional file 4: Figure S3), respectively. In each subgraph, the X-axis represents different feature sets $S_1^1, S_2^1, \ldots, S_9^1, S_{10}^1$, and the Y-axis represents accuracy of each set. For two given feature selection methods M^1 and M^2, if the area under the curve of M^1 is larger than that of M^2, M^1 is better than M^2.

For comparing the performances of the methods using multiple kernels with the method using single kernel, the individual usage and different combination of three kernels, linear kernel $K(x_i, x) = (x_i, x)$, radial basis function kernel $K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2}\right)$ and polynomial kernel $K(x_i, x) = [(x_i, x) + 1]^2$ are conducted. The results of ***mean effectiveness*** and ***max effectiveness*** on three types of datasets are shown (see Additional file 5: Table S2). In Additional file 5: Table S2, the method using different individual kernels affect the results weakly, and the method using multiple kernels has the best results among the majority of the datasets.

In a practical application, the first step can be skipped. However, because of the existing irrelevant features, when only using the second step, the results are not always better than those after removing the irrelevant features, and meanwhile the process has high computational complexity. Considering the computational complexity, we only test the

Table 5 The results of mean effectiveness on miRNA Sequencing (top 10)

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	OSFS	FGM	SMKL-FS
KIDNEY	0.922	0.832	0.987	0.901	0.896	0.893	0.916	**0.994**
BRCA	0.839	0.963	0.979	0.817	0.973	0.893	0.953	*0.990*
LUNG	0.891	0.946	0.979	0.953	0.831	0.945	0.946	**0.980**
HNSC	0.979	0.955	0.991	0.879	0.874	0.920	0.874	**0.994**
LIHC	0.906	0.836	0.911	0.813	0.871	0.789	**0.925**	0.917
PRAD	0.897	0.933	0.930	0.892	0.905	0.794	0.836	*0.946*
STAD	0.855	0.870	0.853	0.790	0.823	0.760	0.827	*0.880*
THCA	0.925	0.901	**0.969**	0.842	0.876	0.878	0.928	0.967
Mean	0.902	0.904	0.950	0.861	0.881	0.859	0.901	**0.958**

performance by only using the second step on miRNA datasets. The results are shown in Additional file 6: Table S3. From the table, we can see that the results of only using the second step are not better than those filtering some features in the first step, and meanwhile using all features the second step has high computational complexity.

Stability of feature space

The stability of feature space generated from a feature selection algorithm reflects the robustness of the method on different samples of the same type of data [34]. For a list of feature sets $S_n^{11}, S_n^{12}, \dots, S_{n^{1k}}$ generated by method M^1 on different samples $\Omega_1, \Omega_2, \dots, \Omega_k$(each Ω is a subset of X) of dataset D and another list of feature sets $S_n^{21}, S_n^{22}, \dots, S_{n^{2k}}$ generated by method M^2 on samples $\Omega_1, \Omega_2, \dots, \Omega_k$. Let $I_1 = \left\{ S_n^{11} \cap S_n^{12} \cap, \dots, \cap S_{n^{1k}} \right\}$, $U_1 = \left\{ S_n^{11} \cup S_n^{12} \cup, \dots, \cup S_{n^{1k}} \right\}$ and $I_2 = \left\{ S_n^{21} \cap S_n^{22} \cap, \dots, \cap S_{n^{2k}} \right\}$, $U_2 = \left\{ S_n^{21} \cup S_n^{22} \cup, \dots, \cup S_{n^{2k}} \right\}$. If $\frac{|I_1| /}{|U_1|}$ is larger than $\frac{|I_2| /}{|U_2|}$, we consider the method M^1 is better than M^2 in ***union stability*** of feature space. For every two samples $\Omega_i, \Omega_j \in \{\Omega_1, \Omega_2, \dots, \Omega_k\}$, let $R_{1^{ij}} = \left| S_{n^{1i}} \cap S_{n^{1j}} \right| / \left| S_{n^{1i}} \cup S_{n^{1j}} \right|$ and $R_{2^{ij}} = \left| S_{n^{2i}} \cap S_{n^{2j}} \right| / \left| S_{n^{2i}} \cup S_{n^{2j}} \right|$, if the average of $R_{1^{ij}}$ is larger than the average of $R_{2^{ij}}$, the method M^1 is better than M^2 in ***independent stability*** of feature space.

In our verification, we set the n of feature sets $S_n^{11}, S_n^{12}, \dots, S_{n^{1k}}$ and feature sets S_n^{21}, $S_n^{22}, \dots, S_{n^{2k}}$ to 100 and use 10-fold cross validation to measure the stability of the feature lists generated by different feature selection methods. Firstly, we randomly choose 90% of the paired samples from each dataset and iterate this process 10 times to obtain 10 different sets for each dataset. Then different feature selection methods are used to select these feature lists. Furthermore, we compute the ***union stability*** and ***independent stability*** according to the process mentioned above.

The results of ***union stability*** on three types of datasets through different methods are shown (See Additional file 7: Table S4). From Additional file 7: Table S4, the ***union stability*** of SMKL-FS is better than those from other methods on most datasets. The results of ***independent stability*** on three types of datasets through different methods are shown in Figs. 1, 2 and 3, respectively. In Figs. 1, 2, 3, the X-axis represents different datasets, and the Y-axis represents ***independent stability***. The ***independent stability*** results of SMKL-FS are better than those from other methods on most datasets.

Similarity with other methods

The similarity between the feature space generated from one feature selection algorithm and the feature lists generated by other methods can be used to estimate the availability of the algorithm. For the feature set S_n^1 generated by method M^1 of dataset D and other feature sets $S_n^2, \dots, S_n^k$ generated by methods $M^2, M^3, \dots, M^k$ of the same dataset D. Let $I_1 = \left| S_n^1 \cap S_n^2 \right|$, $I_2 = \left| \left\{ S_n^1 \cap S_n^3 \right\} \right|$,..., $I_{k-1} = \left| S_n^1 \cap S_n^k \right|$, and $I_{mean} = \frac{1}{k-1} \sum_{j=1^{k-1}} I_j$.

If the I_{mean} of one method is larger than other methods, the method is better than other methods in ***Similarity***.

In our verification, we set n of feature set S_n^1 to 100. Firstly, we select the feature sets $S_n^1, \dots, S_n^6$ on each dataset by SVM-RFE, SVM-RCE, mRMR, IMRelief, SlimPLS and

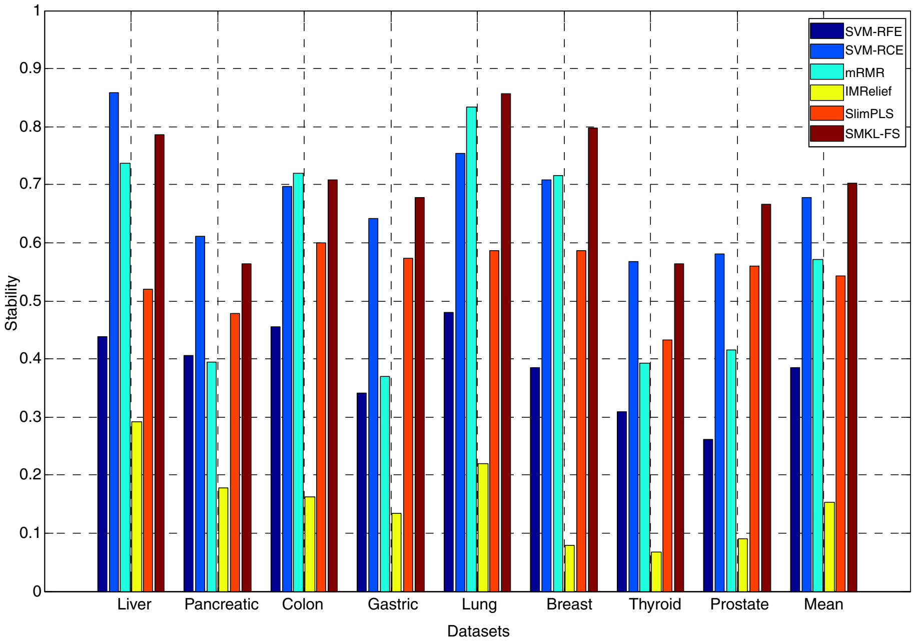

Fig. 1 The results of independent stability on different mRNA microarray datasets

SMKL-FS, respectively. Then, for each feature set generated by every method on one dataset, the value I_{mean} is calculated according to the process mentioned above. The results of ***similarity*** on three types of datasets through different methods are shown in Tables 6, 7 and 8. The ***similarity*** results of SMKL-FS are better than those from other methods on most datasets.

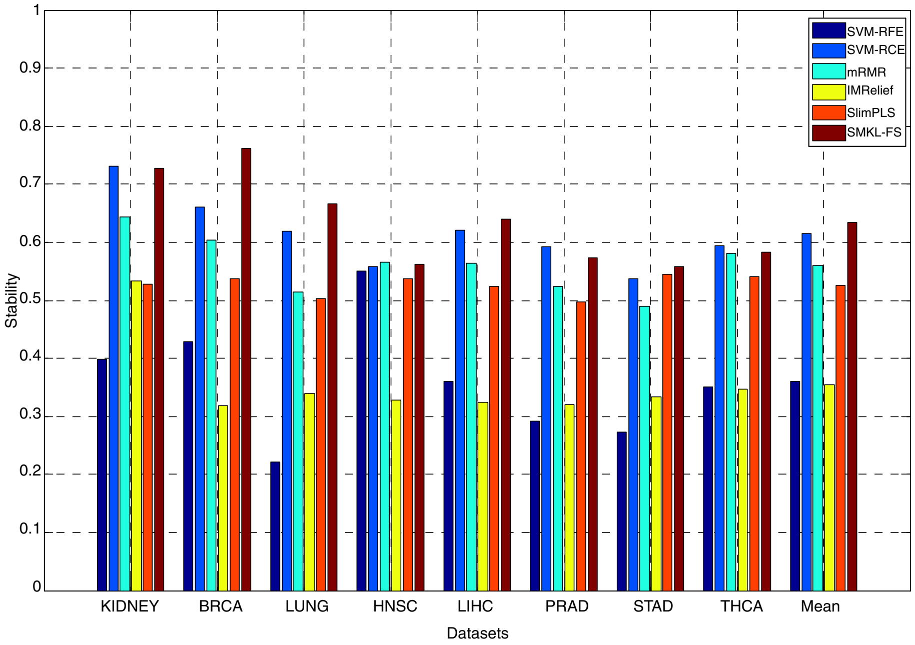

Fig. 2 The results of independent stability on different mRNA Sequencing datasets

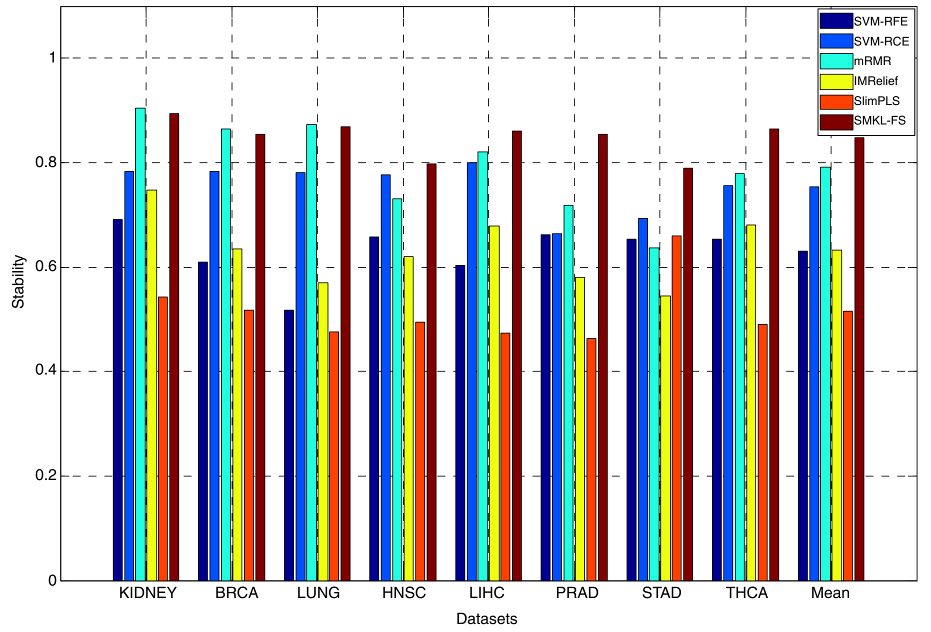

Fig. 3 The results of independent stability on different miRNA Sequencing datasets

Methods

Brief review of SVM

Several supervised learning methods, such as Support Vector Machines (SVMs) can be used to analyze data and recognize patterns by classification and regression analysis. The standard SVM algorithm was proposed by Cortes and Vapnik in 1995 [35]. Given a sample set of data points $G = \left\{ \left(\overrightarrow{x}_i, , y_i \right) \right\}_{i=1}^{n}$, $\overrightarrow{x}_i \in R^m$ and $y_i \in \{+1, -1\}$, the decision rule is:

$$\mathrm{SVM}(x) = \mathrm{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \right) \tag{2}$$

where y_i is the class label of the sample x_i and the summation is taken over all the training samples. α_i is the Lagrange multipliers involved in maximizing the margin of separation of the classes. $K(x_i, x)$ is a kernel which can map the feature space to a high dimensional

Table 6 The results of similarity on mRNA microarray

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	SMKL-FS
Liver	6.33	1.17	**15.83**	1.33	1	15.17
Pancreatic	4.67	0.83	11.17	1.83	3	*16.83*
Lung	3.83	21.83	20.67	0.17	2.17	*23*
Colon	7.17	0.67	19.17	0.67	2.67	*22.67*
Gastric	3.83	0.83	16.00	0.50	3.50	*20.50*
Breast	9.83	32.83	31.83	0	1.67	*33.83*
Thyroid	10.83	29.00	20.17	0	1.67	**29.33**
Prostate	5.50	27.50	20.00	0.50	1.17	*29.17*
Mean	6.50	14.33	19.35	0.63	2.10	*23.81*

Table 7 The results of similarity on mRNA Sequencing

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	SMKL-FS
KIDNEY	1.33	8.00	11.00	2.83	1.67	*12.00*
BRCA	5.67	16.83	14.83	3.67	0.83	*17.83*
LUNG	6.50	23.17	11.50	2.83	0.67	*26.67*
HNSC	1.17	*24.17*	11.67	2.50	1.17	23.00
LIHC	9.50	26.67	17.50	1.33	2.33	*29.33*
PRAD	9.83	26.67	19.17	3.33	0.83	*30.00*
STAD	7.83	**29.67**	15.17	16.67	0.33	29.50
THCA	5.17	14.33	12.50	4.83	0.50	*16.00*
Mean	5.88	21.19	14.17	4.75	1.04	*23.04*

space. There are several popular kernels, such as linear kernel $K(x_i, x) = (x_i, x)$, radial basis function kernels $K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{\sigma}\right)$, homogeneous kernels $K(x_i, x) = (x_i, x)^d$ and inhomogeneous polynomial kernels $K(x_i, x) = [(x_i, x) + 1]^d$. After obtaining the α, we can predict the label of a new data point by the following formula [36]:

$$f(z) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, z) + b \tag{3}$$

and the bias b is defined:

$$b = -\frac{1}{2}\left[\max_{\{i|y_i=-1\}}\left(\sum_{j=1}^{n} \alpha_j y_j K\left(x_i, x_j\right)\right) + \min_{\{i|y_i=+1\}}\left(\sum_{j=1}^{n} \alpha_j y_j K\left(x_i, x_j\right)\right)\right] \tag{4}$$

Multiple kernel learning (MKL)

In recent years, several multiple kernel learning (MKL) methods have been proposed to enhance the interpretability of the decision function and improve performances [23, 24]. A convenient approach of MKL is to construct the kernel $K(x_i, x)$ as a convex combination of basis kernels [23]:

$$K(x_i, x) = \sum_{m=1}^{M} d_m K_m(x_i, x), \qquad \text{with } d_m \geq 0, \ \sum_{m=1}^{M} d_m = 1 \tag{5}$$

where M is the number of multiple kernels. The kernel K_m may be the popular kernels

Table 8 The results of similarity on miRNA Sequencing

Methods	SVM-RFE	SVM-RCE	mRMR	IMRelief	SlimPLS	SMKL-FS
KIDNEY	43.00	33.00	48.50	29.17	28.00	*51.00*
BRCA	39.67	39.33	50.83	25.83	33.00	*52.33*
LUNG	41.50	38.83	50.17	29.50	21.67	*53.33*
HNSC	42.17	38.83	50.50	32.50	22.50	*53.67*
LIHC	42.33	35.50	46.50	24.67	25.17	*47.67*
PRAD	42.33	40.33	53.17	27.00	30.83	*54.33*
STAD	43.50	35.33	48.83	28.67	20.67	*53.33*
THCA	37.33	37.50	47.50	26.50	25.50	*50.83*
Mean	41.48	37.33	49.50	27.98	25.92	*52.06*

mentioned above with different parameters. Each single kernel K_m can either use the full set of training samples or subsets of these samples from different data sources. Then, the problem of the model is transferred to the choice of the weights d_m.

Actually, the standard primal MKL formulation, which just learns from objective consisting of a simple summation of base kernels subjected to mix-norm regularization, is expressed in a functional form as:

$$\min_{f,b,\xi} \frac{1}{2}\left(\sum_m \|f_m\|_{H_m}\right)^2 + C\sum_i \xi_i \;\; s.t.\; y_i\left(\sum_m f_m(x_i) + b\right) \geq 1 - \xi_i,\; \forall i \;\; \xi_i \geq 0\; \forall i \tag{6}$$

where f_m is a function that belongs to corresponding Hilbert space H_m, and each Hilbert space H_m endowed an inner product $\langle \cdot, \cdot \rangle_m$ has a unique kernel K_m.

However, $\|f_m\|_{H_m}$ is not differentiable. When $f_m = 0$, it leads to original objective function, which is not smooth. In this article, we apply SimpleMKL [23] that uses a weighted l_2 norm regularization to calculate the upper bound of the problem through Cauchy-Schwartz inequality. The primal formulation can be replaced as:

$$\begin{aligned} &\min_{f,b,\xi,d} \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|_{H_m}^2 \\ &+ C\sum_i \xi_i s.t.\; y_i\left(\sum_m f_m(x_i) + b\right) \geq 1 - \xi_i,\; \forall i \xi_i \geq 0\; \forall i \sum_m d_m = 1,\; d_m \geq 0,\; \forall \mathrm{m} \end{aligned} \tag{7}$$

And the corresponding dual problem is given as follows

$$\max_{\alpha} \sum_i \alpha_i \text{-} \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_m d_m K_m(x_i, x_j) \quad s.t. \sum_i \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C,\; \forall i \tag{8}$$

where α and C are Lagrange multipliers of the constrains which related to each data point and their tolerable errors separately.

Note that our new dual objective function is convex and differentiable with respect to α. At each iteration, firstly the coefficients keep unchanged, and the value of objective function is optimized. Then, the coefficients are recovered and updated with above dual variables, and this process repeats until convergence.

Feature selection algorithm

Similar to other methods [18, 31], our algorithm also tries to construct an efficient process to select a compact set of features. Combined with the multiple kernel learning (MKL) method mentioned in the above section, we present a two-stage feature selection method. For expression data of a set of features, there are four major feature categories: relevant features, redundant features, irrelevant features and noisy features. For two types of expression data, the relevant features are only a very small part. Most of features are irrelevant features, which will be removed firstly by many feature selection methods for expression data analysis. So, in the first stage of our method, the relevant features are identified by measuring score of each feature using the optimizing process of MKL. If the computational complexity is considered, a small set of relevant features in the first step can be selected. In the second stage, an embedded selection scheme, i.e. the forward selection, is applied to search the subset of compact features from the candidate feature sets obtained in the first stage.

Selecting the relevant feature set

Firstly, we apply MKL to select the relevant feature set. To implement MKL approach, we select the SimpleMKL method in [23] to obtain the coefficient d_m of the kernel combination . SimpleMKL used an iterative gradient descent process to perform an optimization both on the parameters of the SVM (α_i) and the kernel coefficients (d_m). There are several kernels can be used, such as linear kernel $K(x_i, x) = (x_i, x)$, radial basis (RBF) function kernel $K(x_i,x) = \exp\left(-\frac{\|x_i-x\|^2}{2\sigma^2}\right)$ and polynomial kernels $K(x_i, x) = [(x_i, x) + c]^d$.

Then the optimal objective function is defined as follows:

$$J = \min_{d_m} \max_{\alpha} W(\alpha, d_m) \ \text{ such that } \ \sum_{m=1}^{M} d_m = 1 \ , \ d_m \geq 0 \tag{9}$$

Using SimpleMKL, we can obtain the J value for each feature from the total feature set S in the process of optimizing $W(\alpha, d_m)$ via $\min_{d_m} \max_\alpha W(\alpha, d_m)$. To select the relevant feature set, the J list for features list is computed to measure the relevance between features and samples. Finally, we sort the J list in ascend and obtain the ranked features list S_r. Then, the top n^* features are selected and the feature set S_{n^*} is obtained. The process of selecting the relevant feature set is defined (See Additional file 8: Table S5).

Selecting compact feature subsets

An embedded scheme of the sequential forward selection is utilized to search the compact feature subsets from the relevant feature set S_{n^*}. In general, the wrapper approaches convolve with a classifier (e.g., SVM) and the goals are to minimize the classification error of the dependent classifiers. These wrapper approaches can usually obtain low classification error for their dependent classifiers. However, they have high computational complexity and the selected features are less generalization to classifiers [31]. We use the following formula instead of classification error to carry out the embedded process.

$$J_{\mathrm{Z}} = \min_{d_m} \max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_{m}^{M} d_m K_m \left(x_i^{\mathrm{Z}}, x_j^{\mathrm{Z}} \right) \right) \tag{10}$$

where Z is the set containing the selected features, such as $Z = \{f_1, f_2, \ldots, f_n\}$. In this article, the J_{Z} is calculated by using SimpleMKL method [23], which solves the MKL problem through a primitive formulation involving a weighted l2-norm regularization [23].

Then, a forward process is used to to select the subset with r features from S_{n^*} by the incremental manner. And initially, the score of J_0 is set to $+\infty$ and the subset Z is set to empty. We search each feature in the feature subset, such as $f_1, f_2, \ldots, f_n$, and compute the objective functions $J_{f_1}, J_{f_2}, \ldots, J_{f_n}$ using SimpleMKL. The feature f_{i} which generates the largest $\Delta J = J_0 - J_{f_{\mathrm{i}}}$ reduction is appended to Z. Then, the algorithm selects the feature f_{j} which generates the largest ΔJ reduction from the set $\{S_{n^*} - Z\}$ to Z. The process of incremental selection will repeat until $\Delta J \leq 0$ or the given iterations. The process of selecting compact feature subsets is defined (See Additional file 8: Table S6).

Discussion and conclusions

With the development of high-throughput microarray chip and RNA sequencing technology, we can obtain a large number of expression data with different types. The researchers can acquire these data from several public databases, such as GEO, SMD, ArrayExpress and TCGA. However, because the transcriptomics experiments have high cost, most of these data have samples with small size and tens thousands genes or hundreds miRNAs. How to extract informative features from expression data effectively and robustly is a challenging and crucial problem for expression data analysis. Feature selection technique had been widely applied to select a subset of relevant features and eliminate redundant, irrelevant and noisy features.

In general, most feature selection methods can be divided into three categories: filter, wrapper and embedded. Filter methods independent of the classifier, select relevant features only relying on the intrinsic properties of expression data. Filter methods contain two subclasses: univariate and multivariate. Univariate methods are processed by filtering single feature and multivariate methods are used to select features by considering combination of features. The advantages of univariate methods are fast, scalable and independent of the classifier, and the disadvantages of these methods are thoughtlessness of feature dependencies and ignoring the interaction with the classifier. The advantages of multivariate methods contain: feature dependencies, independent of the classifier and better computational complexity than wrapper methods. But the multivariate methods are slower and less scalable than univariate methods. Wrapper approaches, which can be divided into deterministic and randomized types, generate the scores for features and select them based on the classifier. The deterministic methods, which are simple, have less computational complexity and more risk of over fitting than randomized methods. But they are more prone to get a result of local optimum than randomized methods. Embedded approaches, which have lower computational complexity than wrapper methods, select optimal feature subset based on classifier construction in the combined space of feature subsets and hypotheses.

Most of above methods construct the feature selection model on individual expression data simply, and they rarely consider the effectiveness and stability on expression data from different type of expression data. In order to overcome the disadvantages of above methods, a hybrid feature selection method based on multiple kernel learning is proposed. We evaluate performance of method on expression dataset of different types. Except for comparing the classification accuracy with other methods, we also compare the performances of different algorithms through measuring the stability, similarity and consistency. The experimental results show that the proposed method has a satisfactory capability of feature selection for different expression datasets analysis.

The kernel methods and other machine learning methods always have the problem of overfitting, especially in small sample size. And, one of characteristics of gene expression data is high dimensional and small sample size. There are commonly used methodologies to avoid overfitting on machine learning: Regularization, Cross-Validation, Early Stopping and Pruning. The regularization part adds a cost term for bringing in more features with the objective function. Hence, regularization can make the coefficients for many variables to zero and hence avoid the overfitting. Cross validation can identify the methods, which may have poor performance generating by overfitting training on several datasets. The methods of early stopping try to prevent overfitting by controlling the number of iterations on iterative method. Pruning methods, which remove the nodes with little predictive power, are used for several methods based on tree. In this article, we used regularization and sequential forward selection method to decrease the influence of overfitting on small sample size. With the lower price of Mircoarray and RNA sequencing, the samples are more and more obtained from individual experiment, such as the new experiment of RNA sequencing on single-cell, which can handle more than 4000 samples [37]. So, in the future, the influence of overfitting on expression data analysis will be getting smaller and smaller, and machine learning methods and kernel methods will be better used with these data.

Additional files

Additional file 1: Table S1. The results of max effective on mRNA microarray, mRNASeq and miRNASeq datasets. (XLSX 11 kb)

Additional file 2: Figure S1. Classification accuracy of features combination on different mRNA microarray datasets. (PDF 77 kb)

Additional file 3: Figure S2. Classification accuracy of features combination on different mRNASeq datasets. (PDF 61 kb)

Additional file 4: Figure S3. Classification accuracy of features combination on different miRNASeq datasets. (PDF 52 kb)

Additional file 5: Table S2. The results of mean and max effective by different kernel. (XLSX 12 kb)

Additional file 6: Table S3. The results of mean and max effective by using step one. (XLSX 8 kb)

Additional file 7: Table S4. The results of union stability. (XLSX 11 kb)

Additional file 8: Tables S5 and S6. The pseudo code of proposed algorithm. (DOCX 50 kb)

Abbreviations

GEO: Gene Expression Omnibus; miRNA: microRNA; MKL: Multiple kernel learning; MRMR: Minimum Redundancy Maximum Relevancy; mRNA: messenger RNA; RFE: Recursive Feature Extraction; RNA: Ribonucleic acid; SMD: Stanford Microarray Database; SMKL-FS: Simple MKL-Feature Selection; SVM: Support Vector Machine; SVM-RFE: Support Vector Machine-Recursive Feature Elimination; TCGA: The Cancer Genome Atlas

Acknowledgements

We would like to thank Dr. Yan Wang from Jilin University and Dr. Xin Chen of UGA and JLU.

Funding

This work was supported in part by grants from The National Natural Science Foundation of China 61402194, 61472158 and 61572227 and China Postdoctoral Science Foundation 2014 T70291.

Availability of data and materials

Link of data and materials: http://csbl.bmb.uga.edu/ICSB/SMKL-FS/index.html.

Authors' contributions

WD was responsible for the analysis and the draft of the manuscript. ZBC carried out the data analysis and the revision of the manuscript. TCS carried out the revision of the manuscript. YL and YCL participated in the design and the revision of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun, 130012, China. [2]Zhuhai Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Jilin University, Zhuhai, 519041, China. [3]School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun, 130012, China.

Received: 5 November 2015 Accepted: 11 January 2017
Published online: 02 February 2017

References

1. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41(Database issue):D991–5.
2. Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TBK, Wymore F, Zachariah K, Sherlock G, Ball CA. Implementation of GenePattern within the Stanford Microarray Database. Nucleic Acids Res. 2009;37:D898–901.
3. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, et al. ArrayExpress update—trends in database growth and links to data analysis tools. Nucleic Acids Res. 2013;41(Database issue):D987–90.
4. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20.
5. Xu Y, Cui J, Puett D. Cancer Bioinformatics. New York: Springer; 2014: 43.
6. Kim Y, Street WN, Menczer F. Feature Selection in Data Mining. In: Data Mining: Opportunities and Challenges. Hershey: Idea Group Publishing; 2003: 80-105.
7. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
8. Tang Y, Zhang YQ, Huang Z. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans Comput Biol Bioinform. 2007;4(3):365–81.
9. Glaab E, Garibaldi JM, Krasnogor N. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. BMC Bioinformatics. 2009;10:358.
10. Cai H, Ruan P, Ng M, Akutsu T. Feature weight estimation for gene selection: a local hyperlinear learning approach. BMC Bioinformatics. 2014;15:70.
11. Cao ZB, Wang Y, Sun Y, Du W, Liang YC. A novel filter feature selection method for paired microarray expression data analysis. Int J Data Min Bioinform. 2015;12(4):363–86.
12. Mukhopadhyay A, Maulik U. An SVM-wrapped multiobjective evolutionary feature selection approach for identifying cancer-microRNA markers. IEEE Trans Nanobioscience. 2013;12(4):275–81.
13. Maulik U, Chakraborty D. Fuzzy preference based feature selection and semisupervised SVM for cancer classification. IEEE Trans Nanobioscience. 2014;13(2):152–60.
14. Chen Z, Li J, Wei L. A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artif Intell Med. 2007;41(2):161–75.
15. Mao Q, Tsang IW. A feature selection method for multivariate performance measures. IEEE Trans Pattern Anal Mach Intell. 2013;35(9):2051–63.
16. Li Y, Si J, Zhou G, Huang S, Chen S. FREL: A Stable Feature Selection Algorithm. IEEE Trans Neural Netw Learn Syst. 2015;26(7):1388-402
17. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 2014;15:8.
18. Yousef M, Jung S, Showe LC, Showe MK. Recursive cluster elimination (RCE) for classification and feature selection from gene expression data. BMC Bioinformatics. 2007;8:144.
19. Niijima S, Okuno Y. Laplacian linear discriminant analysis approach to unsupervised feature selection. IEEE/ACM Trans Comput Biol Bioinform. 2009;6(4):605–14.
20. Chuang LY, Ke CH, Chang HW, Yang CH. A two-stage feature selection method for gene expression data. OMICS. 2009;13(2):127–37.
21. Mundra PA, Rajapakse JC. SVM-RFE with MRMR filter for gene selection. IEEE Trans Nanobioscience. 2010;9(1):31–7.
22. Du W, Sun Y, Wang Y, Cao ZB, Zhang C, Liang YC. A novel multi-stage feature selection method for microarray expression data analysis. Int J Data Min Bioinform. 2013;7(1):58–77.
23. Rakotomamonjy A, Bach FR, Canu S, Grandvalet Y. SimpleMKL. J Mach Learn Res. 2008;9:2491–521.
24. Gonen M, Alpaydin E. Multiple Kernel Learning Algorithms. J Mach Learn Res. 2011;12:2211–68.
25. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinforma Comput Biol. 2005;3(2):185–205.
26. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1–3):389–422.
27. Gutkin M, Shamir R, Dror G. SlimPLS: a method for feature selection in gene expression-based disease classification. PloS One. 2009;4(7):e6416.
28. Yoon D, Lee EK, Park T. Robust imputation method for missing values in microarray data. BMC Bioinformatics. 2007;8:S6.
29. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30(4):e15.
30. Autio R, Kilpinen S, Saarela M, Kallioniemi O, Hautaniemi S, Astola J. Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. BMC Bioinformatics. 2009;10:S24.
31. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
32. Wu X, Yu K, Ding W, Wang H, Zhu X. Online feature selection with streaming features. IEEE Trans Pattern Anal Mach Intell. 2013;35(5):1178–92.
33. Tan MK, Tsang IW, Wang L. Towards Ultrahigh Dimensional Feature Selection for Big Data. J Mach Learn Res. 2014;15:1371–429.
34. Haury AC, Gestraud P, Vert JP. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. PloS One. 2011;6(12):e28210.
35. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
36. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014;30(6):838–45.
37. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189–96.