Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

A novel feature selection scheme for high-dimensional data sets: four-Staged Feature Selection

Ayça Çakmak Pehlivanlı

To cite this article: Ayça Çakmak Pehlivanlı (2016) A novel feature selection scheme for high-dimensional data sets: four-Staged Feature Selection, Journal of Applied Statistics, 43:6, 1140-1154, DOI: 10.1080/02664763.2015.1092112

To link to this article: https://doi.org/10.1080/02664763.2015.1092112

Journal of Applied Statistics, 2016
Vol. 43, No. 6, 1140–1154, http://dx.doi.org/10.1080/02664763.2015.1092112

Taylor & Francis
Taylor & Francis Group

A novel feature selection scheme for high-dimensional data sets: four-Staged Feature Selection

Ayça Çakmak Pehlivanlı*

Department of Statistics, Mimar Sinan FA University, Cumhuriyet Mh. Silahsor Cd. Bomonti Yerleskesi, Sisli, İstanbul, Turkey

(*Received 8 October 2014; accepted 6 September 2015*)

高维数据集的分类是统计学习和数据挖掘算法面临的一大挑战。为了有效地将分类方法应用于高维数据集，特征选择是学习过程中必不可少的预处理步骤。在本研究中，我们考虑的问题是如何构造一个有效的特征选择和分类方案，用于样本容量小但特征数量多的数据集。针对高维数据分类问题，提出了一种新的特征选择方法——四阶段特征选择方法。该方法首先选择候选特征，并根据不同的度量标准，采用多种滤波方法对候选特征进行筛选，然后分别采用半包装法、联合法和投票法，得到最终的特征子集。为了验证所选特征的有效性，采用了多种统计学习和数据挖掘方法。为了检验所提方法的充分性，我们使用了10种不同的微阵列数据集，因为它们具有数量多、样本容量小的特点。

Classification of high-dimensional data set is a big challenge for statistical learning and data mining algorithms. To effectively apply classification methods to high-dimensional data sets, feature selection is an indispensable pre-processing step of learning process. In this study, we consider the problem of constructing an effective feature selection and classification scheme for data set which has a small number of sample size with a large number of features. A novel feature selection approach, named four-Staged Feature Selection, has been proposed to overcome high-dimensional data classification problem by selecting informative features. The proposed method first selects candidate features with number of filtering methods which are based on different metrics, and then it applies semi-wrapper, union and voting stages, respectively, to obtain final feature subsets. Several statistical learning and data mining methods have been carried out to verify the efficiency of the selected features. In order to test the adequacy of the proposed method, 10 different microarray data sets are employed due to their high number of features and small sample size.

Keywords: feature selection; high-dimensional data; microarray gene expression; statistical filter methods; statistical learning; data mining; classification

Subject Classification Codes: 68T05; 62P10; 92B20; 92D20

1. Introduction

Recently, measuring the expression levels of thousands of genes simultaneously plays a significant role in prediction and clinical diagnosis of diseases. In these microarray experiments, researchers have to deal with a small number of instances which are represented by thousands of genes. Therefore, it is a big challenge for researchers to classify and deal with the

*Email: ayca.pehlivanli@msgsu.edu.tr

high-dimensional data set. Since analysis of this kind of data set involves higher computational complexity and bigger classification error, researchers consider feature selection methods to select the most informative feature subsets. Several studies have shown that to find genes whose expression levels can distinguish different labels plays a crucial role in DNA microarray analysis, because most genes are not relevant for a reliable classification among different classes of the problem [7,16,3,12].

Feature Selection is an essential step to improve classification performance by selecting a proper number of the most relevant and informative properties for analysis. In the field of classification, feature selection methods can be examined in three broad strategies, depending on the way to combine the feature selection search with the construction of the classification model – filter, wrapper and embedded. The filter strategy orders the features without using any learning algorithm and depends on the general characteristics of data [9]. In most cases, a feature score which measures the relevance is calculated; sorted and high-scoring features are considered as the filtered set of features. Afterwards, the learning process is conducted by using this subset of features as the input [26]. For the wrapper approach, the feature selection process involves a learning algorithm and selects the features according to its performance. While wrapper methods embed the induction algorithm to evaluate the feature subsets, filter methods rely on the general structures of data to select features without using any classifier [4]. The third approach of feature selection methods named embedded generally uses machine learning methods for classification, and the search for an optimal subset of features is built into the classifier construction [26].

Although the wrapper method is potentially very time consuming, the criterion can be really optimized for the real problems. In contrast, the filter method is much faster than the wrapper approach and also it is independent of the learning algorithm. On the other hand, while wrapper methods interact with classifier, filter methods ignore the interaction with classifier. Another important drawback of the wrapper method is that it has a possibility of overfitting [26].

In this study, a novel feature selection strategy which can be defined broadly as the combination of filter and wrapper approaches is suggested to overcome the drawbacks listed above. It should be indicated that the wrapper approach used in this scheme is applied without the recursive evaluation step to reduce computational time and complexity. Therefore, in this paper, the 'semi-wrapper' term will be used instead of the 'wrapper' term to avoid confusion. The proposed strategy is introduced in four steps; the first step involves the committee agreement of the several number of filtering methods which are based on different metrics. The aim of the first step is twofold: to make a decision independent of data set and also to reduce the risk of overfitting. The filter step is followed by the semi-wrapper step which applies a classifier to the different number of feature subsets obtained by the first step. The last two steps of the proposed approach involve the union and majority voting to obtain informative and representative final feature subset. This selection procedure is done by using cross-validation to avoid bias selection in features.

The main focus of this paper is to select an effective and more representative feature subset by introducing the novel feature selection scheme called four-Staged Feature Selection (fSFS). It is also designed to reduce variability related to different filtering methods and to eliminate characteristics of data sets.

2. Filter methods

In the scope of this study, four different filtering methods have been used – ReliefF, Effective Ranged based Gene Selection (ERGS), *t*-statistics and Fisher score (FS). Since each method uses a different criterion, they can select different feature subsets and produce different classification success. The brief reviews of these four algorithms are given as follows.

2.1 *Fisher score*

FS [6] tries to find a subset of features such that intra-class distances between data points are as small as possible and inner-class distances between data points are as large as possible. Given the sample set from the jth class $\boldsymbol{X}^j = \{\boldsymbol{x}_1^j, \boldsymbol{x}_2^j, \ldots, \boldsymbol{x}_{n_j}^j\}$ where n_j is the number of samples in the jth class, and let μ_j^i and σ_j^i be the mean and standard deviation of the jth class corresponding to the ith feature, respectively. Fisher criterion, that is, the ratio between class variance to the within class variance, is computed by Equation (1) given below for each ith feature:

$$FS(\boldsymbol{X}^i) = \frac{\sum_{j=1}^{c} n_j(\mu_j^i - \mu^i)^2}{\sum_{j=1}^{c} n_j(\sigma_j^i)^2}, \tag{1}$$

where μ_j^i and σ_j^i denote the mean and standard deviation for the ith feature independent of class information. After calculating FS for each feature, these scores are sorted in descending order to select top m ranked features.

2.2 *t-Statistic*

This filtering method is generally applied to binary class problems and features are selected by t-statistic (TStat) that computed each feature individually by the following formula:

$$\text{TStat}(\boldsymbol{X}^i) = \frac{|\mu_1^i - \mu_2^i|}{\sqrt{((\sigma_1^i)^2/n_1) + ((\sigma_2^i)^2/n_2)}}, \tag{2}$$

where μ_j^i and σ_j^i denote mean and standard deviation of the ith feature for the jth class ($j = 1, 2$), respectively. Similar to FS, features are selected based on the larger values of TStat.

2.3 *ReliefF*

Relief has been introduced by Kira and Rendell in 1992 [13] to assign weights to the features by estimating the quality according to how well their values classify patterns that are close to each other in binary class problems [14]. ReliefF has been proposed as an extension of the basic Relief algorithm to handle multiclass problems, to deal with noisy and incomplete data and to design a more robust approach than the basic version [24]. The ReliefF algorithm selects random pattern T (among m patterns) from class C, finds its k nearest neighbours: k nearest patterns from the same classes, called nearest hit H, and the other k nearest neighbours from different classes, called nearest miss M. The quality estimation w_i for gene i is updated according to its value for pattern T, M and H. It gives less weight if patterns T and H have different values of the feature i, that is, feature i discriminates two patterns with the same class. On the other hand, it increases weight to features that separate the pattern from neighbours of different classes. The algorithm is repeated n times, where n is the user-defined parameter. The update rule that is basically average contribution of all the hits and all the misses is given in the following equation for w_i:

$$w_i = w_i - \frac{\varphi(X_i, T, H)}{nk} + \sum_{C \neq C_T} \left[\frac{\Pr(C)}{1 - \Pr(C_T)}\right] \frac{\varphi(X_i, T, M(C))}{nk}, \tag{3}$$

where function $\varphi(.)$ computes the distance between sample pattern T and nearest hit H or nearest misses $M(C)$, $\boldsymbol{x}_i$ is the ith feature, Pr is the prior probability for each class, $1 - \Pr(C_T)$ represents sum of probabilities for misses' classes.

2.4 *Effective Range based Gene Selection Algorithm*

ERGS has been introduced in 2011 as an efficient feature selection approach based on statistically defined effective range of features for every class [5]. Obtaining the decision boundary of the class by effective range which is uniquely defined by using statistical inference theory is the main principle of the ERGS [10,22]. This novel algorithm computes a weight for each feature. If the decision boundary can easily classify the patterns, that feature is given more weight and also the effective range of the feature does not overlap or has a lesser overlapping area. The ERGS algorithm is true for all distributions because of Chebyshev's inequality and uses the class prior probabilities to calculate effective ranges. In addition to these properties, since there is no iteration process in this one pass algorithm, that is, no need to have a search strategy to implement ERGS, it is faster than any other existing filter methods.

In the original paper [5], effective range is defined by

$$\text{Range}_{ij} = [\text{range}_{ij}^{\text{low}}, \text{range}_{ij}^{\text{up}}] = [\mu_{ij} - (1 - \text{pr}_j)t\sigma_{ij}, \mu_{ij} + (1 - \text{pr}_j)t\sigma_{ij}], \tag{4}$$

where pr_j is the prior probability of the jth class C_j, $(1 - pr_j)$ is the scale factor to regulate the effect of class with high probability and large variance. t is a nonzero real number and its value is defined by using Chebyshev's inequality given in the following formula:

$$P(|X - \mu_{ij}| \geq t\sigma_{ij}) \leq \frac{1}{t^2}. \tag{5}$$

In this study, effective range contains at least two-third of the data object, and then t is calculated as 1.732 as in the original paper [5].

3. four-Staged Feature Selection

In recent years, most of the studies in microarray have focused on feature selections with ensemble approaches, hybrid schemes and decision with consensus [4,15]. Instead of using a single method, it is better to use multiple methods to have unbiased results.

As it is stated earlier in this paper, a novel feature selection strategy which is a basically committee decision is proposed to select an effective and more representative feature subset. Since each filter method that was included in this study uses different strategies and metrics, different feature subsets can be selected from a specific data set, that is, different filter methods select different feature subsets on the same data set. To overcome this high variability, to eliminate characteristics of data sets, to reduce the risk of overfitting and to get results without time consumption are the main concerns of this novel feature selection scheme. It is designed in four main successive stages; *Filter, Semi-Wrapper, Union* and *Voting*. Before processing the stages, k-fold cross-validation (CV) is applied to all data sets to evaluate the stability of the method [2]. During the k-fold CV, the original data set is randomly partitioned into k equal-sized subsamples. Each part is in turn retained as test data, and remaining $k-1$ subsamples are used as training data. The integration of the stages leads to an effective and robust feature selection scheme.

Filter, Semi-Wrapper and *Union* stages are done for all train-test pairs obtained by CV. Therefore these stages are done k times and, for the rest, each of them is defined as fold. The detailed fSFS algorithm is described as Algorithm 1.

Filter. The filter stage which is also given in Algorithm 1 basically uses different filter methods in order to filter out unimportant features and reduces the computational load for the next stages. Instead of using a single filter method, this stage uses several filter methods based on different metrics to reduce the high variability and eliminate characteristics of data sets as stated before.

Algorithm 1 *Four-Staged Feature Selection*

Inputs
$\mathbf{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\} \epsilon \boldsymbol{R}^{\boldsymbol{M}}$: Dataset with all $\boldsymbol{M}$ features $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{\boldsymbol{M}}\}$
$\boldsymbol{K}$: number of folds
Dataset for ***k-fold*** cross-validation: $\boldsymbol{D}_{\boldsymbol{tr}}^{\boldsymbol{k}}, \mathbf{D}_{\mathbf{te}}^{\boldsymbol{k}}, \boldsymbol{k} = 1, \ldots, \boldsymbol{K}$
$\boldsymbol{nbFM}$: number of Filter Methods $(\boldsymbol{FM}_i, \boldsymbol{i} = 1, \ldots, \boldsymbol{nbFM})$
$\boldsymbol{F}_{\boldsymbol{i}}^{\boldsymbol{k}}$: subset with $\boldsymbol{M}^*$ $(\boldsymbol{M}^* \leq \boldsymbol{M}), \boldsymbol{k} = 1, \ldots, \boldsymbol{K}, \boldsymbol{i} = 1, \ldots, \boldsymbol{nbFM}$
$\boldsymbol{M}^*$: user-defined value which is set at the beginning of the algorithm
$\boldsymbol{bestPerf}$: best prediction value
$\boldsymbol{bestFeatNum}$: number of features which produces $\boldsymbol{bestPerf}$
$\boldsymbol{fullSetPerf}^{\boldsymbol{k}}$: performance of a classifier with all features for $\boldsymbol{k}^{\boldsymbol{th}}$ fold, $\boldsymbol{k} = 1, \ldots, \boldsymbol{K}$
$\boldsymbol{CF}_{\boldsymbol{i}}^{\boldsymbol{k}}$: Candidate feature subset obtained by $\boldsymbol{FM}_i$ for $\boldsymbol{k}^{\boldsymbol{th}}$ fold, $\boldsymbol{k} = 1, \ldots, \boldsymbol{K}, \boldsymbol{i} = 1, \ldots, \boldsymbol{nbFM}$
$\boldsymbol{U}^{\boldsymbol{k}}$: Union of features subsets for $\boldsymbol{k}^{\boldsymbol{th}}$ fold $\boldsymbol{k} = 1, \ldots, \boldsymbol{K}$ 第k次折叠的特征子集的并集
Initializations
$\boldsymbol{U}^{\boldsymbol{k}} \leftarrow \{\}, \boldsymbol{CF}_{\boldsymbol{i}}^{\boldsymbol{k}} \leftarrow \{\}, \boldsymbol{k} = 1, \ldots, \boldsymbol{K}, \boldsymbol{i} = 1, \ldots, \boldsymbol{nbFM}$
$k \leftarrow 1$
while$(k \leq K)$
{
carry out a classifier with full features to obtain $\boldsymbol{fullSetPerf}^{\boldsymbol{k}}$
Filter
for each $\boldsymbol{FM}_i; \boldsymbol{i} = 1$ to $\boldsymbol{nbFM}$
perform filter method $\boldsymbol{FM}_i$ on $\boldsymbol{D}_{\boldsymbol{tr}}^{\boldsymbol{k}}$ to calculate importance scores of features
sort features in descending order by means of importance scores obtained by $\boldsymbol{FM}_i$
save the first best $\boldsymbol{M}^*$ features into $\boldsymbol{F}_{\boldsymbol{i}}^{\boldsymbol{k}}$
end for
Semi-wrapper
for each $\boldsymbol{F}_{\boldsymbol{i}}^{\boldsymbol{k}}: \boldsymbol{i} = 1$ to **nbFM** 不同数量的最佳特性子集j（j可以从任意数量开始，j≤M*）
for different number of the best features subsets $\boldsymbol{j}$ (**j** can start from any number, $\boldsymbol{j} \leq \boldsymbol{M}^*$)
apply a classifier on dataset with first best $\boldsymbol{j}$ features selected by $\boldsymbol{FM}_i$ to obtain classification 对FMi选择的第一个最优j特征（前j个特征）的数据集应用分类器进行分类
performance 执行
save prediction performance with corresponding number of features $\boldsymbol{j}$ 保存相应j个数量特征的预测性能
end for
$\boldsymbol{bestPerf} \leftarrow$ max of the prediction performance 最大的预测性能
$\boldsymbol{bestFeatNum} \leftarrow$ number of features which produces $\boldsymbol{bestPerf}$ 产生最佳性能的特性的数量
if $(\boldsymbol{bestPerf} > \boldsymbol{fullSetPerf}^{\boldsymbol{k}})$ 如果最佳性能大于整体性能
store $\boldsymbol{bestFeatNum}$ features to $\boldsymbol{CF}_{\boldsymbol{i}}^{\boldsymbol{k}}$ 将产生最佳性能的特征集存储到CFik中，即候选特征子集
end if
end for
Union
for each $\boldsymbol{i} = 1$ to $\boldsymbol{nbFM}$
append $\boldsymbol{CF}_{\boldsymbol{i}}^{\boldsymbol{k}}$ to $\boldsymbol{U}^{\boldsymbol{k}}$ 将k折叠的好几个候选特征子集取并集放入Uk中
end for
sort $\boldsymbol{U}^{\boldsymbol{k}}$ and remove all duplicates to obtain final subset for $\boldsymbol{k}^{\boldsymbol{th}}$ fold 排序Uk并删除所有重复项，以获得第k次折叠的最终子集
$k \leftarrow k + 1$
}**end while**
Voting
Determine best feature set by majority voting strategy by selecting the most seen features among $\boldsymbol{U}^{\boldsymbol{k}}$s, $\boldsymbol{k} = 1, \ldots, \boldsymbol{K}$ 通过选择Uks中最常见的特性，通过多数投票策略确定最佳特性集，k = 1, …k

看上去和封装差别不大

After performing each filter method on a training data, obtained scores for all features are sorted from the most important (highest score) to the least important (lowest score). Instead of using whole feature set, the proposed method selects first M^* best features from each sorted feature set obtained based on the different metrics (M^* is a user-defined value which is nonzero and less than number of features M). Therefore at the end of this stage, the number of feature subsets is as many as the number of filter methods used in the process for a single fold.

Semi-wrapper。这个半包装器阶段的结构与原始包装器方法不同。在最初的包装器方法中，有一个递归步骤，它从特征空间中消除不重要的特性。每次消除都需要执行一个分类器。在我们的方法中，没有递归步骤，即分类器的应用需要一次遍历，不需要搜索策略。因此，与包装器方法相比，半包装器阶段所需的计算时间要少得多。

Semi-wrapper. The structure of this semi-wrapper stage is different from the original wrapper approach. In the original wrapper approach, there is a recursive step which eliminates insignificant features from the feature space. Each elimination requires carrying out a classifier. In our approach, there is no recursive step, that is, the application of the classifier requires one pass with no search strategy. Thus, the semi-wrapper stage requires significantly less computation time than the wrapper approach.

In this stage, a classifier is carried out with different number of feature subsets that come from the filter stage with M^* best features. As seen in Algorithm 1, classification performances are obtained with different number of the best features subsets (less than M^*). If the best prediction value of these different number of feature subsets is higher than the performance obtained by full feature space, the feature subset which produces the best classification value is saved into a candidate feature subset. This process is done for all feature subsets built by the filter methods that are used in this work. Therefore, the number of candidate subsets depends on the number of filter methods which meets higher prediction performance criterion, that is, there can be *at most* as many as number of filter methods subsets for each fold at the end of this stage.

Union. In the third stage, different numbers of candidate subsets obtained from the second stage are merged into the final subset of each fold and denoted U^k for each fold. At the end of this stage, our feature selection scheme has at most K (number of folds) final subsets. General flow of the union stage is presented in Algorithm 1.

Voting. In the fourth stage, the proposed system decides the final best feature subset by using the majority voting rule on the integrated feature subsets obtained from the third stage. According to the majority voting rule, the features which are seen the most among the $U^{k'}$s, built by union stage, will then be selected as the member of the final best feature set.

Complexity of the fSFS. The computational complexity of the proposed approach basically depends on the classifiers and also filtering methods used in the design. Although the first stage of the algorithm uses several filtering methods, the whole process takes a few seconds, and the complexity of the filtering methods can be omitted. Thus, it is better to consider only the second stage where the classifier method is used to have a general idea about the complexity of the proposed method. The complexity will be roughly $\boldsymbol{clsComp} \times \boldsymbol{nbFS}$ for a single filter; where ***clsComp*** is the complexity of the classifier and ***nbFS*** is the number of the best feature subsets. The number of the filtering methods (***nbFM***, mentioned in Algorithm 1) used in the study should also be considered, therefore the general complexity for a single fold becomes $\boldsymbol{nbFM} \times \boldsymbol{nbFS} \times \boldsymbol{clsComp}$. It should be noted that if the small ***nbFS*** and ***nbFM*** are chosen, the complexity of fSFS is not much more than the complexity of classifier for each fold.

4. Support vector machines

A support vector machine (SVM) introduced by Vapnik is used as the classification method based on statistical learning theory for this study [29]. The main objective of SVM is to find a hyperplane as the decision surface in such a way that the margin between a set of objects of different classes is maximized [11]. To set an optimal hyperplane, maximizing margin which is defined as $2/||\vec{w}\,||$ is equal to minimizing $||\vec{w}^2||/2$, where w is the vector of coefficients. The

SVM is formulated to minimize an error function $J(w)$ given by

$$J(w) = \frac{1}{2}w^{\mathrm{T}}w + C\sum\nolimits_{i}^{\xi}, \tag{6}$$

subject to the constraints:

$$d_i[w^{\mathrm{T}}\varphi(x_i) + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \quad i = 1, \ldots, n, \tag{7}$$

where b is a constant, slack variables ξ_i, $i = 1, \ldots, n$, are parameters to allow classification error and they are used with parameter C which controls the trade-off between the size of margin and the number of nonseparable points. x_i's are independent variables represented by class labels d_i for each training sample i. $\varphi(.)$ is a kernel function to construct nonlinear decision boundary. It transforms input data into higher dimensional feature space by simply adding an additional dimension by means of kernel functions.

In this study, exponential kernel given by Equation (8) has been used as the kernel function. The optimal value of trade-off parameter C and kernel parameter σ has been determined by grid search and CV. As a result of the series run, C and σ were chosen as 1.5 and 3, respectively.

$$\varphi(x, x_i) = \exp\left(-\frac{||x - x_i||}{2\sigma^2}\right). \tag{8}$$

5. Model evaluation 模型预测

The results are quantified with the following statistical parameters: Sensitivity (Sens), Specificity (Spec) and Accuracy (Acc). These parameters were used to measure the accuracy of positive prediction, for example, the percentage of tumour predicted correctly, negative prediction, for example, the percentage of nontumor predicted correctly, and the overall accuracy of the model, respectively. The metrics were calculated with the following equations:

$$\text{Sens} = TP/(TP + FN), \tag{9}$$

$$\text{Spec} = TN/(TN + FP), \tag{10}$$

$$\text{Acc} = (TP + TN)/(TP + TN + FP + FN), \tag{11}$$

where ***TP* (true positive)** is the number of correctly classified tumours (cancer, failures, etc.); ***FP* (false positive)** the number of nontumor instances incorrectly classified as tumour; ***TN* (true negative)** the number of correctly classified nontumors and ***FN* (false negative)** the number of tumours incorrectly classified as nontumor.

通常，总精度Acc总是用来衡量模型[17]的预测能力。所有给定的测度都受到相对类频率的部分影响，不足以进行无偏估计。因此，还计算了曲线下的概率过剩(ProbEx)和接收机工作特性面积(AUC)，以进行比较。

In general, the overall accuracy Acc is always used to measure the prediction power of a model [17]. All given measures are partly affected by the relative class frequency, and they are not enough to unbiased evaluation. Therefore, the probability excess (*ProbEx*) and Receiver Operating Characteristic Area Under the Curve (AUC) values are also calculated for comparison.

The *ProbEx* that varies between 0 and 1 is calculated by (Sens + Spec − 1). While accuracy is affected by the relative frequencies of the two classes, the *ProbEx* value is independent of the relative class frequency in the data set [31]. 在0和1之间变化的ProbEx由(Sens + Spec - 1)计算。虽然精度受两个类的相对频率影响，但ProbEx值与数据集中[31]的相对类频率无关。

The ROC curve plots true positive rate (sensitivity) against the false positive rate (1 − specificity) for a classifier as a function of decision cut-off (e.g. probability of tumour class to be classified as nontumor class). The AUC is especially useful as a clue for robustness of algorithm rather than for the comparison of different algorithms [31].

ROC曲线将分类器的真正率(敏感性)与假正率(1 -特异度)作为决策截止函数绘制出来(例如肿瘤类别被分类为非肿瘤类别的概率)。AUC作为算法鲁棒性的线索，比不同算法[31]的比较更有用。

Table 1. Details of experimental data sets I.

Data set	Samples	Genes	Explanation
Central nervous system (CNS) 中枢神经系统	60	7129	This data set was designed to analyse the outcome of the certain treatment. After the treatment, there are survivors and failures [21] 2 classes: survivors vs failures
Colon tumour (Colon) 结肠肿瘤	62	2000	The data were collected from tumour biopsies of colon cancer patients and nontumor biopsies of the same patients [1] 2 classes: tumour vs normal
Non-small cell lung carcinoma in female non-smokers (feLung) 女性非吸烟者的非小细胞肺癌 (feLung)	120	54,675	This data set was designed as genome-wide screening of transcriptional modulation in non-smoking female lung cancer to analyse the molecular signature of non-smoking female lung cancer in Taiwan [18] 2 classes: lung cancer vs normal
Prostate 前列腺	136	12,600	2 classes: prostate tumour vs nontumor [27]
ALL-AML Leukaemia (Leukaemia) ALL-AML白血病	72	7129	This data set consists of bone marrow samples with ALL and AML [7] 2 classes: ALL vs AML

该数据集旨在分析特定治疗的结果。经过治疗，有幸存者和失败者[21]
数据收集自结肠癌患者的肿瘤活检和相同患者[1]的非肿瘤活检
本研究以全基因组筛选非吸烟女性肺癌的转录调控为目的，分析台湾地区非吸烟女性肺癌的分子特征
前列腺肿瘤和非肿瘤
这个数据集由骨髓样本组成 ALL和AML [7]

为了检验提出的特征选择方案的性能，对10组基因表达谱数据集进行了实验。数据集具有少量的实例，这些实例可用大量的基因来解释。本研究中使用的每个微阵列数据集用N × M矩阵表示，其中N为实例数，M为基因数。所有的数据集都是二分类，而且每个类的分布通常不均匀。女性非吸烟者的PDAC、多发性骨髓瘤及骨病变、黑色素瘤转移及非小细胞肺癌数据集从Gene Expression Omnibus data set Browser[19]下载，其余数据集从Kent Ridge Biomedical data set Repository[23]下载。数据集的实验细节见表1和表2。

6. Microarray data sets

In order to test the performance of the proposed feature selection scheme, the experiments are carried out on 10 data sets of gene expression profiles. The data sets have a small number of instances which are explained with the high number of genes. Each microarray data set used in this study is represented by the $N \times M$ matrix, where N is the number of instances and M is the number of genes. All the data sets are binary classes, and each class is generally unevenly distributed. PDAC, Multiple myeloma and bone lesions, Melanoma Metastases and Non-small cell lung carcinoma in female non-smokers data sets are downloaded from Gene Expression Omnibus Data set Browser [19] and rest are obtained from Kent Ridge Biomedical Data set Repository [23]. Experimental details of the data sets are given in Tables 1 and 2.

7. Experimental results

实验结果

This section presents evaluation of a newly proposed fSFS scheme on 10 microarray data sets. In the experiments, both the feature selection procedure and the statistics of classification performance are evaluated by 10-fold CV. Each data set is partitioned into 10 equal-sized data sets. Nine subsamples are used as training data set, selecting optimal features, and remaining subsample is retained for testing the model. In order to apply the proposed four-staged method, different filter methods should be chosen for the first stage. In this experiment, four different filtering methods, Relieff, ERGS, TStat and FS are involved by considering their different metrics. After applying these methods to sort the features from the most important to least important, first $M^* = 100$ best features have been chosen among each sorted set. This process has been applied to all training sets obtained by cross-validation. At the end of this stage, each fold has four different feature subsets, $F_{FS}, F_{\text{TStat}}, F_{\text{Relieff}}, F_{\text{ERGS}}$, with 100 top-ranked genes. In the second stage, training has been carried out for different subsets of these selected genes from 10 to 100 by increasing 10 for each fold. Candidate subsets of each fold are chosen according to their classification performance. At the end of the second stage, there are *at most* four subsets (because of the number of filter methods used in this experiment) with different number of features for each fold. By integrated these subsets, *at most* 10 subsets (because of the number of folds used

本节介绍了一种新的基于10个微阵列数据集的fSFS方案的评估。在实验中，用10折交叉验证对特征选择过程和分类性能的统计量进行了评价。每个数据集被划分为10个大小相等的数据集。选取9个子样本作为训练数据集，选取最优特征，保留剩余子样本对模型进行测试。为了应用所提出的四阶段滤波方法，第一阶段应选择不同的滤波方法。在本实验中，考虑了四种不同的滤波方法Relieff、ERGS、TStat和FS的不同指标，采用了四种不同的滤波方法。将这些方法应用于从最重要到最不重要的特性排序之后，首先在每个排序集中选择M *=100个最佳特性。该过程已应用于交叉验证得到的所有训练集。在这个阶段结束时，每个折叠有四个不同的特征子集，FFS、FTStat、FRelieff、FERGS，共有100个顶级基因。在第二阶段，对这些选定基因的不同子集进行训练，从10个增加到100个，每次增加10个。根据每个折叠的分类性能选择候选子集。在第二阶段的末尾，最多有四个子集(由于本实验中使用的过滤方法的数量)，每个折叠具有不同数量的特征。对这四个子集进行合并（四个子集取并集），在合并阶段结束时最多可以得到10个子集(因为CV使用的折叠数)。每个数据集的最终特征子集由多数投票方案实现，该方案利用来自折叠的10个子集作为所提出的特征选择过程的最后阶段。

Table 2. Details of experimental data sets II.

Data set	Samples	Genes	Explanation
Lung cancer (Lung)	181	12,533	This data set has malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung [8] 2 classes: MPM vs ADCA
Melanoma Metastases (Melanoma)	83	22,283	This data set represents melanoma clinical samples with primary melanomas and melanoma metastases [30] 2 classes: primary melanoma vs melanoma metastases
Multiple myeloma and bone lesions (Myeloma)	173	12,625	Gene expression in bone marrow plasma cells of multiple myeloma patients with and without bone lesions are in this data set [28] 2 classes: without bone lesions vs with bone lesions
Ovarian	253	15,154	This data set was designed to identify proteomic patterns in serum that distinguish ovarian cancer from non-cancer. The women who have a high risk of ovarian cancer due to family or personal history of cancer were included into data set [20] 2 classes: cancer vs normal
Pancreatic ductal adenocarcinoma (PDAC)	90	28,869	Analysis of 45 matching pairs of PDACtumour and adjacent nontumor tissue [32] 2 classes: pancreatic tumour vs nontumor

for CV) are obtained at the end of the union stage. The final feature subset for each data set is achieved by the majority voting scheme that utilized among the subsets coming from the folds as the last stage of the proposed feature selection process.

为了更清楚地解释所提出的系统，以CNS数据集为例给出了详细的结果。CNS数据集包含7129个基因(特征)的60个样本。在应用第一阶段之前，将CNS数据集划分为10个相等的部分，以便对性能进行较为全面的比较。对于每个折叠，分别对分割的数据集应用四种不同的滤波方法，每一种滤波方法将7129个基因减少到100个排名前位的基因。由于每个滤波方法为每个褶选择自己的排序特征集，因此在第一个阶段结束时，会得到40个不同的排序特征集(4种滤波方法×10个褶)。首先对每个具有完整特征的折叠进行训练，然后应用于四种滤波方法得到的100个最佳特征的不同子集。表3给出了第二阶段部分折叠的滤波方法的分类精度(为了节省位置，只给出了前两次和最后两次折叠)。

7.1 *An example: detailed results of CNS data set*

一个例子:CNS数据集的详细结果

In order to explain the proposed system more clearly, detailed results of the CNS data set are given as an example. The CNS data set contains 60 samples of 7129 genes (features). Before applying the first stage, CNS data set is divided into 10 equal parts in order to obtain a relatively comprehensive comparison on the performances. For each fold, four different filtering methods are applied on divided data sets individually and 7129 genes are reduced to 100 top-ranked genes by each filtering method. Since each filter method selects its own sorted feature sets for each fold, 40 different sorted features are obtained (4 filter methods × 10 folds) at the end of this first stage. Training is first carried out for each fold with full features and then applied to the different subsets of 100 best features obtained by four filtering methods. Table 3 presents the classification accuracies of filtering methods for some folds as a result of the second stage (in order to save places only first two and last two folds are presented).

According to Table 3, one can analyse all the individual results that are obtained by different number of best features and overall accuracies with full gene sets. Since all individual results of the first fold are greater than or equal to its overall accuracy with full genes set, four feature subsets are selected from fold 1. Thirty features from the results of Relieff yield 67%, 10 features from the results of ERGS with 83%, 10 from *t*-statistics with 67% and 50 from the results of FS with 83% meeting lower error criterion required for stage 2. Since all results of Relieff and FS are smaller than or equal to overall accuracy with full gene set for fold 2, no features are selected from their results. Therefore, at the end of the second stage, while four feature subsets come from fold 1, two feature subsets come from fold 2, no feature subset is obtained from fold 3, etc. The union stage merges these subsets to get a single subset for each fold and 85, 60, 0, 10, 75, 50, 60,

根据表3，我们可以分析所有由不同数量的最佳特征得到的单个结果以及完整基因集的整体准确性。由于第一个折叠的所有个体结果都大于或等于完整基因集的总体精度，因此从折叠1中选择4个特征子集。Relieff结果中30个特征的符合率为67%，ERGS结果中10个特征的符合率为83%，t-statistics结果中10个特征的符合率为67%，FS结果中50个特征的符合率为83%，满足阶段2所需的较低的误差标准。由于折叠2中Relieff和FS的所有结果都小于或等于完整基因集的总体精度，因此它们的结果中没有选择任何特征。因此，在第二阶段结束时，虽然有4个特征子集来自于折叠1,2个特征子集来自于折叠2，但是没有一个特征子集来自于折叠3，等等。并集阶段将这些子集合并得到每个褶的一个子集，得到CNS数据集的85、60、0、10、75、50、60、44、49和126个基因子集，分别为1-10个褶。最后对这些子集应用多数投票规则，得到CNS数据集的最终特征集，共有27个基因。同样的程序也适用于本研究中使用的所有微阵列数据集。

Table 3. Results of the CNS data set for 10 folds[a].

Fold #	All	Method	10	20	30	40	50	60	70	80	90	100
1	0.50	Relieff	0.50	0.50	**0.67**	0.50	0.33	0.33	0.33	0.50	0.50	0.50
		ERGS	**0.83**	0.67	0.67	0.83	0.83	0.83	0.83	0.67	0.67	0.67
		Tstat	**0.67**	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
		FScore	0.33	0.50	0.67	0.67	**0.83**	0.83	0.83	0.83	0.83	0.83
2	0.67	Relieff	0.50	0.50	0.33	0.33	0.50	0.33	0.33	0.50	0.67	0.67
		ERGS	0.67	**0.83**	0.83	0.67	0.67	0.67	0.67	0.67	0.67	0.67
		Tstat	0.67	0.67	0.67	**0.83**	0.83	0.83	0.67	0.67	0.83	0.83
		FScore	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
...				...		...				...		
9	0.67	Relieff	**1.00**	0.50	0.83	0.83	0.83	0.83	0.67	0.67	0.67	0.67
		ERGS	0.83	0.83	**1.00**	1.00	1.00	1.00	0.83	0.83	0.83	0.67
		Tstat	0.67	0.50	0.50	0.67	0.67	0.67	0.33	0.33	0.33	0.33
		FScore	0.67	**0.83**	0.67	0.67	0.83	0.83	0.83	0.83	0.83	0.83
10	0.67	Relieff	0.50	0.67	0.67	0.50	0.50	0.50	0.50	0.67	**0.83**	0.83
		ERGS	0.67	0.67	0.67	**0.83**	0.83	0.67	0.83	0.83	0.83	0.83
		Tstat	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
		FScore	0.50	0.50	0.50	0.67	0.67	0.67	0.67	0.67	0.67	0.67

Note: [a]To save places, it is only 4 of the 10 folds are displayed in Table 3.

Table 4. 10-fold CV classification accuracies of CNS, Colon, feLung with SVM.

Data set	Algorithm	Acc	ProbEx	Spec	Sens	Auc	# of genes
CNS	Proposed method	**0.867**	**0.667**	0.950	**0.717**	**0.938**	27
	With full data set	0.647	0.000	**1.000**	0.000	0.638	7129
	Relieff	0.700	0.150	**1.000**	0.150	0.469	
	ERGS	0.767	0.367	0.950	0.417	0.653	
	Tstat	0.633	0.083	0.850	0.233	0.544	
	FScore	0.683	0.192	0.875	0.317	0.592	
Colon	Proposed method	**0.887**	**0.767**	0.900	**0.867**	**0.888**	27
	With full data set	0.640	0.000	**1.000**	0.000	0.796	2000
	Relieff	0.817	0.55	0.90	0.65	0.89	
	ERGS	0.776	0.51	0.83	0.68	0.86	
	Tstat	0.852	0.67	0.90	0.77	0.83	
	FScore	0.867	0.70	0.90	0.80	0.85	
feLung	Proposed method	**0.967**	**0.933**	0.967	0.967	0.994	24
	With full data set	0.482	0.000	0.800	0.200	**1.000**	54,675
	Relieff	0.950	0.90	0.95	0.95	0.99	
	ERGS	0.958	0.92	**0.97**	0.95	0.97	
	Tstat	**0.967**	0.93	**0.97**	**0.97**	0.99	
	FScore	**0.967**	0.93	**0.97**	**0.97**	0.99	

44, 49 and 126 genes subsets of CNS data set are obtained, respectively, 1–10 folds. At the last stage, the majority voting rule is applied to these subsets and the final feature set with 27 genes is obtained for the CNS data set. The same procedures are applied to all microarray data sets used in this study.

7.2 *Comparison of the proposed method and individual filter methods*

将提出的方法与单独的滤波方法进行了比较

Several data mining and statistical learning methods are applied to the final feature subset to compare our approach to individual results of filtering methods and also to compare the results achieved by full data sets. In order to save places, only results obtained by SVM are given in detail in Tables 4–6.

最后的特征子集采用了多种数据挖掘和统计学习方法，以比较我们的方法与单个过滤方法的结果，并比较完整数据集的结果。为了节省位置，表4-6中只给出SVM得到的结果。

Table 5. 10-fold CV classification accuracies of Leukaemia, Lung, Melanoma and Myeloma with SVM.

Data set	Algorithm	Acc	ProbEx	Spec	Sens	Auc	# of genes
Leukaemia	Proposed method	0.959	0.897	**0.967**	0.930	0.988	23
	With full data set	0.648	0.000	0.000	**1.000**	0.848	7129
	Relieff	**0.967**	**0.93**	0.93	1.00	0.993	
	ERGS	0.957	0.91	0.93	0.98	0.980	
	Tstat	0.957	0.89	0.92	0.98	**1.000**	
	FScore	0.957	0.91	0.93	0.98	0.987	
Lung	Proposed method	**1.000**	0.985	0.990	0.905	0.997	27
	With full data set	0.823	0.000	**1.000**	0.000	1.000	12,533
	Relieff	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	
	ERGS	0.989	0.960	0.993	0.967	**1.000**	
	Tstat	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	
	FScore	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	
Melanoma	Proposed method	0.914	0.837	0.903	**0.933**	0.967	22
	With full data set	0.627	0.000	**1.000**	0.000	**1.000**	22,283
	Relieff	**0.938**	**0.860**	0.960	0.900	0.953	
	ERGS	0.915	0.820	0.920	0.900	0.955	
	Tstat	0.914	0.823	0.923	0.900	0.928	
	FScore	0.914	0.837	0.903	**0.933**	0.928	
Myeloma	Proposed method	**0.844**	**0.391**	0.949	**0.442**	**0.893**	33
	With full data set	0.789	0.000	**1.000**	0.000	1.000	12,625
	Relieff	0.787	− 0.008	0.992	0.000	0.604	
	ERGS	0.800	0.165	0.952	0.213	0.630	
	Tstat	0.807	0.171	0.968	0.204	0.678	
	FScore	0.814	0.190	0.968	0.222	0.652	

Table 6. 10-fold CV classification accuracies of Prostate, Ovarian and PDAC with SVM.

Data set	Algorithm	Acc	ProbEx	Spec	Sens	Auc	# of genes
Ovarian	Proposed method	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	47
	With full data set	0.640	0.000	**1.000**	0.000	0.982	15,154
	Relieff	0.988	0.972	0.994	0.978	**1.000**	
	ERGS	0.976	0.949	0.982	0.967	0.996	
	Tstat	0.992	0.978	**1.000**	0.978	0.999	
	FScore	0.996	0.989	**1.000**	0.989	**1.000**	
PDAC	Proposed method	**0.900**	0.795	0.860	**0.935**	**0.960**	37
	With full data set	0.489	0.000	0.200	0.800	0.975	28,869
	Relieff	0.844	0.695	0.785	0.910	0.890	
	ERGS	**0.900**	**0.800**	**0.865**	**0.935**	0.930	
	Tstat	0.867	0.725	0.810	0.915	0.880	
	FScore	0.867	0.730	0.840	0.890	0.935	
Prostate	Proposed method	**0.868**	**0.731**	0.868	**0.863**	**0.957**	66
	With full data set	0.566	0.000	**1.000**	0.000	0.662	12,600
	Relieff	0.836	0.671	0.845	0.827	0.922	
	ERGS	0.846	0.693	0.843	0.850	0.935	
	Tstat	0.824	0.629	0.882	0.747	0.892	
	FScore	0.845	0.681	0.868	0.813	0.921	

在表4-6中，我们可以将所提出的四阶段法得到的结果与所提出方案中涉及的各个滤波器分别得到的结果以及全数据集的结果在灵敏度(Sens)、特异性(Spec)、准确度(Acc)、概率过剩(probEx)和AUC方面进行比较。表中的粗体表示最佳结果。

In Tables 4–6, one can compare the results obtained by the proposed four-Staged method with the results obtained by each of the filters involved in the proposed scheme separately and the results of full data sets in terms of terms of the Sensitivity (*Sens*), Specificity (*Spec*), Accuracy (*Acc*), Probability Excess (*probEx*) and AUC. The bold fonts in the tables represent the best results.

Regarding individual results obtained by filter methods in Tables 4–6, it can be seen that there is no filtering method that always produces the best results. For the Lung data set, all filter

对于表4-6中滤波方法得到的单个结果，可以看出没有一种滤波方法总是产生最好的结果。对于Lung数据集，除ERGS外，所有的滤波方法都能产生100%的分类精度。在剩下的9个数据集中，FS对6个数据集工作得很好，ERGS和Relieff对3个数据集工作得很好，t-statistics对2个数据集工作得很好。这种变化来自于每个过滤器方法使用的不同度量。为了克服这种变化，提出了一种fSFS方案，将滤波方法与自身的子集相结合。

methods produce 100% classification accuracy except ERGS. Among the remaining nine data sets, it seems that FS works well for six data sets, both ERGS and Relieff work adequately three data sets, and *t*-statistics works well for two data sets. This variation arises from the different metrics used by each filter method. The fSFS scheme is proposed to overcome this variation by merging the filter methods with their own subsets.

根据表4-6所示的结果，提出的四种过滤器的委员会决策在按Acc比较单个结果时，获得了10个数据集中8个数据集的最佳结果。另一方面，与所提出的分类结果和单独的结果相比，具有完整特征的分类结果最差。因此，其余的分析将在建议的方法和表4-6中给出的单个结果的最大值之间进行。

In the light of the results given in Tables 4–6, the proposed committee decision of four filters achieves the best results of 8 out of 10 data sets when comparing individual results in terms of Acc. On the other hand, classifications with full features yield the worst results comparing the proposed and individual results. Therefore, the rest of the analyses will be done among the proposed approach and the maximum of the individual results that are given in Tables 4–6.

The results of lung cancer given in Table 5 are the same as the results of the proposed method and the results achieved by individual filtering methods except ERGS. In the same table, one can analyse that the proposed method yields a decrease in success rate when comparing the best individual accuracy produced by Relieff of Leukaemia and Melanoma, however both the proposed method and Relieff yield very close AUC value which is the metric for robustness of the algorithm.

表5所示肺癌的结果与所提出方法的检测结果一致，也与除ERGS外的单个滤波方法的结果一致。在同一个表中，我们可以分析，当比较缓解白血病和黑色素瘤的最佳个体准确率时，提出的方法的成功率会降低，然而，提出的方法与Relieff的AUC值非常接近，AUC值是算法鲁棒性的度量指标。

According to the results reported by Tables 4–6, the proposed committee decision in this research, formed by four filtering methods, is the one which produces the best results. In general, considering average of all data sets, it is observed approximately by 2.9%, 8.4% and 6.7% increments in the accuracy, probability excess and AUC, respectively. The increments in probability excess values indicate that the proposed method yields a more balanced result for two classes and also the increments in AUC values demonstrate that the proposed method is more robust than the individual filtering methods used in this study.

根据表4-6所报告的结果，本研究中提出的四种滤波方法组成的委员会决策是产生最佳结果的决策。：总的来说，考虑到所有数据集的平均值，在准确率、概率过剩和AUC上分别约增加了2.9%、8.4%和6.7%。概率过剩值的增量表明，该方法对两个类的结果更加均衡，AUC值的增量表明，该方法比本研究中使用的单个滤波方法具有更强的鲁棒性。

7.3 *Comparison with other studies*
与其他研究的比较

In order to address the completeness of the study, the results of some of the data sets are compared with those published by Bolon-Canedo *et al*. [4] and Ruiz *et al*. [25]. Bolon-Canedo *et al*. proposed a new framework for feature selection consisting of an ensemble of filters and classifiers. They used C4.5, Naive Bayes and IB1 as the classifiers. Ruiz *et al*. proposed a wrapper method named BIRS (Best Incremental Ranked Subset). They also used the same classifiers to test their approach. The performances over the data sets Colon and Leukaemia are compared with results of Ruiz *et al*. and Colon, Leukaemia, CNS, Prostate, Lung and Ovarian are contrasted with those provided by Bolon-Canedo *et al*. [4]. According to results presented in Table 7, it can be concluded that the proposed fSFS scheme resulted in the lowest average classification error for both IB1 and Naive Bayes. The results obtained by fSFS with C4.5 and IB1 are very close to results obtained by Ensemble methods with C4.5 and IB1 classifiers for Prostate data set. In the same way, fSFS-IB1 and Ensemble-IB1 produced almost the same classification error for both Lung and Ovarian data sets.

为了说明研究的完整性，将部分数据集的结果与Bolon-Canedo等人发表的[4]和Ruiz等人发表的[25]进行了比较。Bolon-Canedo等人提出了一种新的特征选择框架，该框架由一组过滤器和分类器组成。他们使用C4.5、Naive Bayes和IB1作为分类器。Ruiz等人提出了一种名为BIRS（Best Incremental rank子集）的包装器方法。他们还使用相同的分类器来测试他们的方法。将结肠和白血病数据集的性能与Ruiz等人的结果进行比较，并将结肠、白血病、中枢神经系统、前列腺、肺和卵巢数据集与Bolon-Canedo等人提供的[4]进行比较。由表7所示的结果可以看出，所提出的fSFS方案对IB1和朴素贝叶斯的平均分类误差最小。使用C4.5和IB1的fSFS得到的结果与使用C4.5和IB1前列腺数据集分类器的集成方法得到的结果非常接近。同样，fSFS-IB1和Ensemble-IB1对肺和卵巢数据集产生的分类错误几乎相同。

7.4 *Effects of different number of folds*
不同折叠次数的效果

As stated previously, our approach uses the committee decision of the folds. Experiments with different number of folds are done to see the effect of the folds. Table 8 compares the results obtained by 3-fold, 5-fold and 10-fold experiments. In order to compare results, five different well-known classification methods which are C4.5, Naive Bayes, IB1, Random Forest and SVM are used. It should be cleared here that only the feature selection process with fSFS was carried out with different number of folds; classification results given in Table 8 are still achieved by 10-fold CV. To reduce variability, 10 runs of CV were performed using different partitions, and the results displayed in Table 8 are averaged over the runs.

如前所述，我们的方法使用折叠委员会的决策。通过对不同折痕数的实验，研究了不同折痕数的折痕效果。表8为3折、5折和10折实验结果对比。为了比较结果，我们使用了5种不同的著名分类方法，分别是C4.5、朴素贝叶斯、IB1、随机森林和SVM。这里需要澄清的是，只有使用fSFS的特征选择过程是在不同的折叠次数下进行的；表8中给出的分类结果仍然是通过10折交叉验证。为了减少可变性，使用不同的分区执行了10次CV运行，表8中显示的结果在运行期间取平均值。

我认为是分别用3折、5折、10折得出特征子集，对最终的特征子集用分类器检验（10折交叉验证）。

Table 7. Comparison of proposed methods with Bolon-Canedo *et al*. [4] and Ruiz *et al*. [25] – 10-fold CV classification errors obtained for C4.5, Naive Bayes and IB1.

Classifier	Algorithm	Colon	Leukaemia	CNS	Lung	Ovarian	Prostate	Avg. error
C4.5	With full data set	17.742	18.056	51.667	4.972	7.115	16.177	19.288
	fSFS – proposed method	20.381	15.143	**32.50**	3.629	2.806	11.832	14.382
	Ensemble [4]	**13.100**	**11.960**	36.670	**2.750**	**1.200**	**11.810**	**12.915**
	BIRS [25]	19.355	15.278	–	–	–	–	–
Naive Bayes	With full data set	43.548	**1.389**	43.333	2.210	7.115	45.588	23.864
	fSFS – proposed method	13.905	4.196	**20.667**	1.096	2.171	**38.978**	**13.502**
	Ensemble [4]	16.190	4.110	30.000	**0.000**	**0.800**	41.870	15.495
	BIRS [25]	**12.900**	4.167	–	–	–	–	–
IB1	With full data set	19.355	13.889	41.667	4.972	5.929	17.647	17.243
	fSFS – proposed method	**18.595**	9.054	**16.00**	1.143	0.355	**12.481**	**9.605**
	Ensemble [4]	19.050	**5.540**	36.670	**1.110**	**0.000**	12.530	12.483
	BIRS [25]	20.240	6.960	–	–	–	–	–

Table 8. Ten-fold CV classification accuracies obtained for several classifiers with 3-fold, 5-fold and 10-fold fSFS.

Classifiers	k-fold fSFS	CNS	Colon	feLung	Leuk.	Lung	Melan.	Myelo.	Ovarian	Pdac	Prost.	avg
C4.5	**3**	0.635	0.789	**0.918**	0.842	**0.965**	**0.861**	0.775	0.980	0.808	0.869	0.844
	5	0.643	0.778	0.913	0.848	0.960	0.848	**0.800**	**0.981**	0.802	0.853	0.843
	10	**0.675**	**0.796**	0.913	**0.849**	0.964	0.858	0.766	0.972	**0.811**	**0.882**	**0.849**
Naive Bayes	**3**	**0.847**	**0.873**	**0.964**	**0.965**	0.992	0.927	0.767	**0.980**	0.876	0.607	0.880
	5	0.778	0.851	0.961	0.960	0.994	0.940	0.823	0.971	0.891	0.602	0.877
	10	0.793	0.861	0.959	0.958	0.989	0.939	**0.852**	0.978	**0.893**	**0.610**	**0.883**
IB1	**3**	0.755	0.803	0.938	**0.951**	**0.989**	**0.905**	0.778	**1.000**	**0.834**	0.842	0.880
	5	0.782	0.752	0.952	0.949	0.982	0.902	0.796	0.993	0.806	0.854	0.877
	10	**0.840**	**0.814**	**0.953**	0.909	**0.989**	0.895	**0.800**	0.996	0.820	**0.873**	**0.889**
Random Forest	**3**	0.775	0.860	**0.963**	0.970	**0.990**	0.920	**0.858**	0.993	0.867	**0.930**	0.913
	5	**0.797**	0.850	0.965	**0.978**	0.989	0.925	0.831	**0.994**	0.869	0.903	0.910
	10	0.783	**0.865**	0.960	0.976	0.989	**0.931**	**0.858**	0.988	**0.872**	0.917	**0.914**
SVM	**3**	0.840	0.871	0.965	**0.975**	**0.999**	**0.953**	0.843	**1.000**	0.869	**0.864**	0.918
	5	**0.867**	0.871	0.965	0.972	0.997	0.939	0.859	0.999	**0.898**	0.842	0.921
	10	0.860	**0.880**	**0.970**	0.963	0.995	0.941	**0.865**	0.999	0.896	**0.864**	**0.923**

Note: Leuk., Leukaemia; Melan., Melanoma; Myelo., Myeloma; Prost., Prostate; SVM, support vector machines.

Table 9. Number of features selected by fSFS with different folds.

k-fold fSFS	CNS	Colon	feLung	Leukaemia	Lung	Melanoma	Myeloma	Ovarian	Pdac	Prostate	avg
3	78	**20**	34	47	65	82	26	57	**36**	65	51
5	42	36	31	33	37	26	**22**	52	52	**43**	37
10	**27**	27	**24**	**23**	**27**	**22**	33	**47**	37	66	**33**

Regarding the features selected by using fSFS with 3-, 5- and 10-fold, as obviously seen in Table 8, the best average classification accuracies among the results of five different classifiers are achieved by the features selected by 10-fold fSFS. These results should be evaluated with the number of features obtained by different folds given in Table 9. In the light of the results given in Tables 8–9, fSFS formed by the higher number of folds is the one which achieves the best average classification accuracies with the smallest average number of features for all five classifiers.

对于使用3倍、5倍和10倍fSFS选择的特性，如表8所示，利用10倍fSFS选择的特征，在五种分类器的结果中获得了最佳的平均分类精度。这些结果应该用表9中给出的不同折叠得到的特征数来评估。根据表8-9的结果，由较高的折叠次数形成的fSFS是所有五个分类器中平均特征个数最小，平均分类精度最好的分类器。

近年来，许多统计学习和数据挖掘算法被提出用于高维数据集的分类。由于这些数据集具有特征数量多，样本量小的特点，降低搜索空间复杂度是一个挑战，也是必不可少的。本文提出了一种新的特征选择方案统计方法fSFS，用于选择相关的、信息量大的、数量较少的特征进行分类。为了提出该方法，我们选择了10种不同的高维微阵列基因表达数据集，因为它们具有数量多、样本量小的特点。
该方法分为过滤、半包装、联合和投票四个阶段；它首先缩小特征空间进行快速搜索，然后使用分类器进行评估，并将候选特征合并到每个折叠的单个子集中，然后根据多数投票规则确定最终的特征集。一旦得到最终的特征子集，就使用五种著名的数据挖掘方法来测试该方法的性能。所有分类结果均采用10折交叉验证得到。在fSFS的特征选择过程中，也进行了不同次数的折叠实验，但本文给出的所有比较和结果都是基于10倍CV得到的结果。

8. Conclusions

结论

Recently, many statistical learning and data mining algorithms have been proposed for classification of high-dimensional data sets. Reducing the search space complexity is a challenge and indispensable for these data sets due to the high number of features with small sample size. In this paper, a novel statistical approach of feature selection scheme, fSFS has been proposed in order to select relevant, informative and small number of features for classifying. In order to present the proposed approach, 10 different high-dimensional microarray gene expression data sets were chosen due to their high number of features and small sample size.

The proposed approach was conducted in four stages named filtering, semi-wrapper, union and voting; it first narrows down the feature space for the fast search, then evaluates with a classifier, and merges the candidate features into a single subset for each fold followed by deciding the final feature set by the majority voting rule. Once the final feature subset was obtained, five well-known data mining methods were used to test performance of the proposed method. All classification results have been obtained by using 10-fold CV method. During the feature selection process by fSFS, different number of folds was also experimented, but all comparisons and results given in this paper are based on the results obtained 10-fold CV.

实验结果表明，fSFS方法取得的结果优于8 / 10个数据集的单个特征选择方法取得的结果，也优于不选择特征的整个数据集取得的结果。并在6个数据集上与其他研究者的结果进行了比较，发现我们的方法比其他两个分类器的平均分类精度好。

The experimental results present that the results achieved by the fSFS method are better than the results achieved by individual feature selection methods 8 over 10 data sets and the results of whole data sets without feature selections over all data sets. The results were also compared with the results achieved by the other studies with six data sets and observed that our approach performs well with two classifiers in terms of average classification accuracies.

Since this novel approach can be conducted with any filtering and classification methods, it could be considered as a general and flexible framework for the feature selection from high-dimensional data sets. Although this flexible structure is proposed with only two class problems, it can be easily used with the multiclass problems either by choosing filtering and learning methods that are applicable to multiclass problems or by simply constructing one-to-others approach.

In conclusion, most of the feature selection methods generally select the top-ranked features according to their individual distinguishing performance. Since each uses different metrics and statistical approaches, their performances can vary among the different data sets. The proposed method introduced in this paper is designed to reduce this variability in addition to robustness and effectiveness with high classification performance.

由于这种新方法可以使用任何过滤和分类方法，它是一种通用的、灵活的特征选择框架。虽然这种灵活的结构只提出了两类问题，它可以通过选择适用于多类问题的过滤和学习方法，或者通过简单地构造一对其他方法，轻松地用于多类问题。

Disclosure statement

综上所述，大多数的特征选择方法一般都是根据各自的区分性能来选择排名靠前的特征。由于每种方法使用不同的度量标准和统计方法，它们的性能在不同的数据集之间可能有所不同。本文提出的方法在保证鲁棒性和有效性的同时，还具有较高的分类性能。

No potential conflict of interest was reported by the author.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide arrays*, Proc. Nat. Acad. Sci. U.S.A. 96 (1999), pp. 6745–6750.

[2] C. Ambroise and G.J. McLachlan, *Selection bias in gene extraction on the basis of microarray gene-expression data*, Proc. Nat. Acad. Sci. U.S.A. 99 (2002), pp. 6562–6566.

[3] P. Baldi and A. Long, *A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes*, Bioinformatics 17 (2001), pp. 509–519.

[4] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, *An ensemble of filters and classifiers for microarray data classification*, Pattern Recogn. 45 (2011), pp. 531–539.

[5] B. Chandra and M. Gupta, *An efficient statistical feature selection approach for classification of gene expression data*, J. Biomed. Inform. 44 (2011), pp. 529–535.

[6] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed., John Wiley and Sons, New York, 2001.

[7] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, J. Sci. 286 (1999), pp. 531–537.

[8] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, and R. Bueno, *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and Mesothelioma*, Cancer Res. 62 (2002), pp. 4963–4967.
[9] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction. Foundations and Applications*, Springer, New York, 2006.
[10] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer-Verlag, Berlin Heidelberg, 2007.
[11] S. Haykin, *Neural Networks and Learning Machines*, Pearson, Upper Saddle River, NJ, 2009.
[12] P. Jafari and F. Azuaje, *An assessment of recently published gene expression data analyses: Reporting experimental design and statistical factors*, BMC Med. Inform. Decis. Making 6 (2006), p. 27.
[13] K. Kira and L.A. Rendell, *A Practical Approach to Feature Selection.* Proceedings of the 9th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1992, pp. 249–256.
[14] I. Kononenk, *Estimating Features: Analysis and Extension of RELIEF.* Proceedings of the 6th European Conference on Machine Learning, Springer-Verlag New York, Inc., Secaucus, NJ, 1994, pp. 171–182.
[15] C. Lee and Y. Leu, *A novel hybrid feature selection method for microarray data analysis*, Appl. Soft Comput. 11 (2011), pp. 208–213.
[16] L. Li, T.A. Darden, C.R. Weinberg, A.J. Levine, and L.G. Pedersen, *Gene assessment and sample classification for gene expression data using a genetic algorithm/ k-nearest neighbour method*, Comb. Chem. High T. Scr. 4 (2001), pp. 727–739.
[17] Q. Li, A. Bender, J. Pei, and L. Lai, *A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification*, J. Chem. Inf. Model. 47 (2007), pp. 1776–1786.
[18] T.P. Lu, M.H. Tsai, J.M. Lee, C.P. Hsu, P.C. Chen, C.W. Lin, J.Y. Shih, P.C. Yang, C.K. Hsiao, L.C. Lai, and E.Y. Chuang, *Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women*, Cancer Epidem. Biomar. 19 (2010), pp. 2590–2597.
[19] NCBI, *Gene Expression Omnibus Dataset Browser* (*GEO Dataset Browser*). Available at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser/ (Accessed June 2013)
[20] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, and L.A. Liotta, *Use of proteomic patterns in serum to identify ovarian cancer*, Lancet 359 (2002), pp. 572–577.
[21] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T.P., S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, *Prediction of central nervous system embryonal tumour outcome based on gene expression*, Lett. Nature. 415 (2002), pp. 436–442.
[22] C.R. Rao, *Linear Statistical Inference and its Application*, John Wiley and Sons, New York, 1965.
[23] K. Ridge, *Kent Ridge Bio-Medical Dataset*. Available at http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html (Accessed May 2012).
[24] M. Robnik-Sikonja and I. Kononenko, *Theoretical and empirical analysis of relief and rReliefF*, Mach. Learn. J. 53 (2003), pp. 23–69.
[25] R. Ruiz, J.C. Riquelme, and J.S. Aguliar-Ruiz, *Incremental wrapper-based gene selection from microarray data for cancer classification*, J. Pattern Recogn. 39 (2006), pp. 2383–2392.
[26] Y. Saeys, I. Inza, and P. Larranaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics 23 (2007), pp. 2507–2517.
[27] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers, *Gene expression correlates of clinical prostate cancer behavior*, Cancer Cell 1 (2002), pp. 203–209.
[28] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J.D. Jr Shaughnessy, *The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma*, N. Engl. J. Med. 349 (2003), pp. 2483–2494.
[29] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
[30] L. Xu, S.S. Shen, Y. Hoshida, A. Subramanian, K. Ross, J.P. Brunet, S.N. Wagner, S. Ramaswamy, J.P. Mesirov, and R.O. Hynes. *Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases*, Mol. Cancer Res. 6 (2008), pp. 760–769.
[31] Z.R. Yang, R. Thomson, P. McNeil, and R.M. Esnouf, *RONN: The Bio-basis function neural network technique applied to the detection of natively disordered regions in proteins*, Bioinformatics 21 (2005), pp. 3369–3376.
[32] G. Zhang, A. Schetter, P. He, N. Funamizu, J. Gaedcke, B.M. Ghadimi, T. Ried, R. Hassan, H.G. Yfantis, D.H. Lee, C. Lacy, A. Maitra, N. Hanna, H.R. Alexander, and S.P. Hussain, *DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma*, PLoS One. 7 (2012), e31507.