Knowledge-Based Systems 95 (2016) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features

Qifeng Zhou[a], Hao Zhou[a], Tao Li[b,c,*]

[a] School of Aerospace Engineering, Automation Department, Xiamen University, Xiamen, 361005, China
[b] School of Computing and Information Sciences, Florida International University, Miami, FL, 33199, United States
[c] School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210046, China

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 9 September 2015
Accepted 11 November 2015
Available online 11 December 2015

Keywords:
Cost sensitive
Feature selection
Random forest

ABSTRACT

Feature selection aims to select a small subset of informative features that contain most of the information related to a given task. Existing feature selection methods often assume that all the features have the same cost. However, in many real world applications, different features may have different costs (e.g., different tests a patient might take in medical diagnosis). Ignoring the feature cost may produce good feature subsets in theory but they can not be used in practice. In this paper, we propose a random forest-based feature selection algorithm that incorporates the feature cost into the base decision tree construction process to produce low-cost feature subsets. In particular, when constructing a base tree, a feature is randomly selected with a probability inversely proportional to its associated cost. We evaluate the proposed method on a number of UCI datasets and apply it to a medical diagnosis problem where the real feature costs are estimated by experts. The experimental results demonstrate that our feature-cost-sensitive random forest (FCS-RF) is able to select a low-cost subset of informative features and achieves better performance than other state-of-art feature selection methods in real-world problems.

1. Introduction

The feature selection (FS) problem has been studied by the statistics and machine learning communities for many years. Its main theme is to select a small subset of informative features that best discriminate the data objects of different classes [1]. In many data analysis tasks, feature selection is an important and frequently used dimensionality reduction technique and is often considered as a critical step in data pre-processing. In addition, feature selection can significantly improve the understandability of the machine learning models and often help build models with better generalization [2,32]. As a result, in many situations, finding a good subset of features is an important problem in its own right.

Feature cost, literally meaning the cost consumed in acquiring a features value, is a special case of various cost types in machine learning and data mining [9–11]. It may involve different factors such as money, time, and implementation difficulty. In many real world applications, however, different features may have different costs and the difference may be sufficiently large to influence the result of feature selection. Take medical diagnosis as an example, we collected a dataset regarding hepatitis from a local hospital, the costs of most features from this dataset are estimated by experts, in terms of price, time, implementation difficulty, and side effect (i.e., medical tests, their detailed descriptions are given in Section 5). As seen in Table 1, different features may vary considerably across the costs. If a classifier is constructed with many expensive features, to predict new samples, it needs to pay a high cost and the classifier can be thought as lack of practicability. In such cases, it would be better to use a feature subset with an acceptable classification performance but a much lower cost. This kind of feature selection considering the cost is called cost-sensitive feature selection whose aim is to select both low-cost and informative features.

Existing feature selection techniques can be approximately divided into three categories: embedded, filter, and wrapper methods [1,3–8,31]. However, all these three categories of FS methods usually only focus on the features' distinguishing ability or their contribution to classification, and ignore their cost. One simple case of the filter methods Relief-F, for example, is to seek the subsets of features whose pairwise correlation is low. Support vector machine recursive feature elimination (SVM-RFE) [5], a more sophisticated embedded feature selection method, operate by eliminating the least weighted feature in the construction of the SVM model recursively. These traditional methods, in effect, assume all the features have the same cost.

[*] Corresponding author at: School of Computing and Information Sciences, Florida International University, United States. Tel.: +1 305 348 6036; fax: +1 305 348 3549.
E-mail address: taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.knosys.2015.11.010
0950-7051/© 2015 Elsevier B.V. All rights reserved.

Table 1
Some medical tests and their costs in hepatitis treatment.

Cost / test	G	S	ALT	AST	HBV-DNA	Gene-type	MTCT
Price(¥)	100–200	100–200	89	89	140	390	–
Time(day)	3–7	3–7	1	1	2-3	4-5	–
Difficulty(0–10)	7	7	3	3	4	5	2
Side effect(0–10)	3–4	3–4	1	1	1	1	0

A branch specialized to handle cost-related problems, in data mining, is cost-sensitive learning [6,13]. The taxonomy of different types of costs is summarized by Turney [12] and the most common types of cost are the misclassification cost and the feature cost [14,15]. Both of them often arise in practical applications. For example, in medical diagnosis, the feature cost (or the test cost), is based on the fact that the medical tests, whose results the physician will refer to in diagnosing a patient, vary in the running time, the expense, and the side effect etc. The physician needs to estimate the tradeoff between the test effect and its associated cost before deciding which tests the patient should take. The test cost is actually one special case of the feature cost (i.e., the cost of acquiring one feature's value). In the following, we use these two terms (feature cost and test cost) interchangeably and name the process of acquiring a feature's values as having or taking a test.

Compared with the misclassification cost, the feature cost has been studied much less. In reality, feature cost is not only difficult to quantify, but also can have more complicated scenarios, e.g., some feature cost is variable, and different features costs may be connected. These scenarios are summarized by Min and Liu [13] who construct a hierarchical model covering six possible test-cost-sensitive decision systems.

Basically, there are two strategies to reduce the feature cost for a data mining task. The first one is to work out some principles or rules about how to utilize the features for a new test instance. This strategy is especially suitable for a small amount of test instances with many missing values: for every new instance, whose feature values are partially unknown, the strategy decides which feature should be used or tested (if its value is unknown). Note that different instances may vary on unknown features. To predict a new instance, a tailored classification model are needed (usually unsophisticated, like trees). This strategy, which effectively focuses on the classification process rather than the regular feature selection, finds its way in many applications, e.g., the decision tree-based methods, including ICET [12], minimal cost tree [14,15]; and other applications including Markov Decision Process [17,18], and some general test tactics [14,16]. The second strategy for reducing feature cost is to search for both informative and low-cost feature subsets. Models trained with such a feature subset can retain its structure in the test stage, and therefore, is usually fast and applicable for a great amount of test instances. This strategy, in essence, is an improvement on the ordinary feature selection. However, it is a more complex global optimization process, as it takes cost into consideration. Compared against the first strategy, the second strategy as well as its potential benefit has been rarely studied to date.

In order to obtain low-cost subsets of informative features, one straightforward solution is to take a two-step approach: first performing feature selection in regular way, then further analyzing the generated feature subsets or rank based on the costs. However, since these two steps are conducted separately, the interaction between the features discriminative ability and the costs will be neglected. As a result, the ultimate outcome will depend primarily on the regular feature selection, and the goal of "cheap" often cannot be satisfied. In addition, since the second analysis step is conducted manually based on experts experience, there is much uncertainty for the results. Another solution to seeking good and cheap feature subsets is making use of SVM [19], as it can assign weights to the features (this characteristic is already used in the classical feature selection method SVM-RFE [8]). By incorporating the cost factor into the SVM optimization processing, it is possible to approximately re-calibrate the weights for the features. The expanded SVM model can be viewed in [20]. The main limitation of this method is that the outcomes are very sensitive to the model parameters (for themselves, they are also difficult to set) thus weakening its practicability.

To overcome the aforementioned limitations, in this work, we propose a random forest-based cost-sensitive feature selection method named feature-cost-sensitive random forest (FCS-RF). FCS-RF can sort the features based on their comprehensive performance both on the distinguishing ability and costs. The top ranked features will be selected into the final feature subset first. Specifically, the FCS-RF incorporates the feature cost information into the base decision tree construction process. When constructing a base tree, a feature is selected with a probability inversely proportional to its associated cost, instead of being selected randomly. By means of the underlying mechanism of random forest, the importance of all features is calculated and a feature rank can be obtained considering both the feature cost and the distinguishing ability.

The contributions of our work are summarized as follows:

(1) We propose a cost-sensitive feature selection method FCS-RF, which overcomes the limitation of two-step cost-sensitive feature selection methods. FCS-RF incorporates both the distinguishing ability (or quality) of features and their costs as criteria into one optimization process. Therefore, FCS-RF is an approximate global optimization method which can consider the correlations among the features;

(2) We perform a series of empirical evaluations on benchmark datasets to demonstrate the effectiveness of FCS-RF. Compared with the commonly used feature selection approaches, FCS-RF can reduce the feature costs while maintaining a comparable classification performance;

(3) We apply FCS-RF to a real-world medical diagnosis to find cheap and good feature subset for interferon-α (IFN-α) treatment of hepatitis B virus(HBV) . The evaluating results on more than 300 real cases of patients demonstrate the effectiveness of FCS-RF.

The rest of the paper is organized as follows. Section 2 gives the formulation of our problem. Section 3 presents the feature-cost-sensitive random forest algorithm. Section 4 describes how to produce a cost-sensitive feature rank using the improved random forest. Section 5 shows the experiments and the results analysis. Finally, Section 6 makes conclusions and discusses the future work.

2. Problem formulation

A basic feature-cost-sensitive decision system can be summarized according to [15] as

$$S = (U, F, D, V_a | a \in F \cup D, I_a | a \in F \cup D, c^*), \tag{1}$$

where U is a finite set of objects called the universe, F is the set of features, D is the set of class variables (decisions), V_a is the set of values for each $a \in F \cup D$, I_a: $U \to V_a$ is an information function for each $a \in F \cup D$, $c^* : F \to R^+ \cup 0$ is the feature cost function. Considering some records of patients $\{x_1\ x_2\ \dots\ x_r\}$ with each record containing the information of age, gender and gene type, then $U = \{x_1\ x_2\ \dots\ x_r\}$, $F = \{age, gender, genetype\}$, $D = \{response, noresponse\}$. It is noticed that if c^* is empty, the system will ignore the feature cost (i.e., assuming all features have the same cost). This paper focuses on a general situation when the features are parallel and obtained independently, indicating that a constant value can be assigned to each feature as the feature cost. Let n be the number of features, we can use the following vector to signify the costs of all features.

$$C = [c_1 \; c_2 \; \ldots \; c_n]. \tag{2}$$

By feature selection, a subset $B \subseteq F$ can be obtained, and the total feature cost is

$$c_B = \sum_{b \in B} c_b. \tag{3}$$

Because the features are obtained independently, the total cost of a subset equals the sum of costs of individual features. Without loss of generality, we define the minimal cost value as 1. That means for features that are cost-free, their cost will be set as 1. It should be noted that, it is the relative value between different feature cost values that affects the process and the outcomes of cost-sensitive feature selection. Clearly, minimizing the feature cost is not the objective of feature selection and the quality of features (i.e. their contribution to classification) must be considered as well. For simplicity, the classification error, actually a special form of the misclassification cost, is introduced and defined as Eq.(4)

$$err = \frac{\sum_{m}^{i=1} e(predict\_label(x_i), true\_label(x_i))}{m}, \tag{4}$$

where m is the number of test samples and

$$e(x, y) = \begin{cases} 0 & \text{if} \quad x = y; \\ 1 & \text{if} \quad x \neq y. \end{cases} \tag{5}$$

Then the objective of feature selection is to seek a feature subset B^* that could minimize

$$TC = err(B^*) + \alpha \sum_{b \in B^*} c_b, \tag{6}$$

where $err(B^*)$ denotes the classification error when using subset B^*, and α denotes a tradeoff between the classification error and the feature cost. It is feasible to enumerate all possible subsets if the number of original features is small. However, when the number of features is large, the brute force method will become prohibitive. A key issue to solve Eq. (6) is to balance the two terms in the right side. Our strategy is to minimize the feature cost on the premise that classification error is acceptable. Such an assumption is reasonable given that (1) some redundant features exist and to remove them does not affect the accuracy of classifier while reducing the total feature cost; (2) some expensive features of high quality (distinguishing ability) can be replaced with the low-cost features of similar quality, thus also reducing the total cost.

3. Feature-cost-sensitive random forests

A feature-cost-sensitive tree is a variant of the ordinary decision tree with the modified splitting criterion that takes the feature cost into consideration. Typical examples can be seen in [21–23], but a single tree usually cannot match the random forest in terms of the generalization capability. In addition, a single tree is rarely used for feature selection whereas the random forest, an ensemble of trees, is more appropriate to perform feature selection [24]. In our work, we extend the random forest by incorporating a probability vector into the tree construction process.

3.1. Random forest

A decision tree is grown in a recursive fashion by partitioning the training samples into purer subsets successively. At each recursive step, the samples in a node are split into several subsets (child nodes) based on the feature that could minimize the impurity of the child nodes. The splitting process will stop when the node is pure or its impurity is lower than a threshold, and in that case, the node (named leaf node) will be assigned a class label.

Random forest (RF) is an ensemble of decision trees [25]. RF has a wide range of applications because of its good stability and generalization [26–28]. The typical construction process of RF consists of the following procedures. First, bagging [29] is adopted on the training dataset to produce many subsets (with differences). Then each subset is used to construct a decision tree. In the tree growth, the splitting on each node depends on the feature selected from a group of candidates that are randomly chosen from all features. Without pruning (i.e., all leaf nodes are pure), all trees grow fully and each of them functions as a base classifier. Finally, all these tree classifiers are integrated. There are two important random characteristics in growing a random forest. One is randomly sampling, and the other is randomly generating the node splitting candidates. The diversity between the trees caused by randomness is crucial to the performance of the forest.

3.2. Feature-cost-sensitive random forest algorithm

We incorporate probability into the tree construction process when choosing the group of candidate features and let the probability of a feature being selected inversely proportional to its cost. The algorithm of growing a feature-cost-sensitive random forest (FCS-RF) is summarized in the pseudo-codes. Note that *mtry* is the number of the candidate features in each split, *ntree* is the size of the random forest and ΔI_k is the information gain using the kth feature.

Algorithm 1 Feature-cost-sensitive random forest.

Input: M samples with the features $F = [f_1 \; f_2 \; \ldots \; f_n]$, feature cost vector $C = [c_1 \; c_2 \; \ldots \; c_n]$,
Output: feature-cost-sensitive random forest

```
INITIALIZATION
Set $mtry$, $ntree$.
for $i = 1$ to $n$ do
    $\gamma_i = \frac{1}{c_i}$
    $p_i = \frac{\gamma_i}{\gamma_1 + \gamma_2 + \cdots + \gamma_n}$
end for
The probability vector $P = [p_1 \; p_2 \; \ldots \; p_n]$
PROCESS
for $j = 1$ to $ntree$ do
    $node = 1$;
    while the node is impure do
        the candidate features group $G = [\ ]$;
        while $length(G) \leq mtry$ do
            select one feature $f \in F$ based on $P$
            if $f \notin G$ then
                $G = [G \; f]$
            end if
        end while
        compute $\Delta I_k, k \in \{l | f_l \in G\}$ and $f^* = f_{argmax \; \Delta I_k}$
        node splits using $f^*$
        $node{+}{+}$
    end while
end for
```

The probability vector is generated based on the cost vector and we require the probability of a feature being selected inversely proportional to its cost. It can be easily shown that the expected cost EX of the feature selected in probability is smaller than the expected cost EY of the feature selected randomly .

Table 2
Feature rank variation on dataset Wine.

Cost	Feature rank												
1	**10**	**7**	**13**	12	1	6	5	11	2	3	4	9	8
2	12	1	**10**	**13**	**7**	5	11	6	2	4	3	9	8
3	12	1	**10**	5	11	**7**	6	**13**	2	9	4	3	8
4	1	12	11	5	6	**10**	2	**13**	**7**	9	3	4	8
5	12	1	5	11	6	2	**10**	9	4	**7**	3	**13**	8
10	12	1	5	11	6	2	9	4	3	**10**	**7**	**13**	8
20	1	12	11	5	6	2	3	9	4	8	**10**	**13**	**7**
40	1	12	5	11	6	2	9	3	4	8	**7**	**13**	**10**

Proposition 1. *$EX \le EY$. Note that we have*

$$EX = \sum_{i=1}^{n} p_i c_i. \tag{7}$$

$$EY = \frac{1}{n} \sum_{i=1}^{n} c_i. \tag{8}$$

By substituting p_i with the representation from the algorithm, we can simplify EX and get

$$EX = \frac{n}{\frac{1}{c_1} + \frac{1}{c_2} + \cdots + \frac{1}{c_n}}. \tag{9}$$

Then to prove $EX \le EY$ is equivalent to prove the following inequality

$$\left(\frac{1}{c_1} + \frac{1}{c_2} + \cdots + \frac{1}{c_n}\right)(c_1 + c_2 + \cdots + c_n) \ge n^2. \tag{10}$$

which can be proved using mathematical induction given $c_i \ge 1$. The above proposition shows that, by incorporating the probability, the random forest becomes feature cost sensitive. But some randomness among the trees still remains, which is indispensable to preserve the accuracy. It should be noted that if all the cost of features are equal, the FCS-RF will degenerate to the ordinary RF.

4. Searching for low-cost and informative features

Random forest provides the mechanism to calculate the importance of all features, leading to a feature rank. By permuting the values of one feature for all samples, the increment of OOB (out of bag, i.e. the samples not used in growing a tree) error or the decrease of OOB accuracy can be used to calculate the importance score of that feature. Specifically, it consists of the following steps.

(1) Put the OOB samples on the *j*th tree and calculate its accuracy A_j^0, $j = 1, 2, \ldots, ntree$;
(2) For the *i*th feature, permute its values for the OOB samples and test these samples on the *j*th tree to get its accuracy A_j^i, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, ntree$;
(3) The importance score of the *i*th feature is measured by

$$s_i = \frac{1}{ntree} \sum_{j=1}^{ntree} \left(A_j^0 - A_j^i\right). \tag{11}$$

(4) Sort the scores in descending order and generate a list $\{s_{d_1}, s_{d_2}, \ldots s_{d_n}, \}$. Then the vector $\{d_1, d_2, \ldots, d_n\}$ is the feature rank.

The entry in the rank signifies the index of the feature and top ranked features have high importance.

For feature-cost-sensitive random forest, we have incorporated the feature cost into the construction of random forest, making the high-cost features less likely to be selected and the low-cost features with larger chances of being selected. In this way, the importance score of a feature is explicitly influenced by feature cost and its distinguishing ability. As a result, the top ranked features in the rank produced through FCS-RF have larger chance to be both low-cost and informative.

Although the variation of OOB accuracy is quantitatively measurable, the effect of the probability in tree growth is hard to quantify. In this work, we will validate the effectiveness of the proposed algorithm via a series of experiments. In particular, we expect the proposed method works well in the situations where the redundant features exist, or the expensive features have cheaper substitutes with similar quality.

5. The experiments

The experiments are conducted on a number of datasets, including 10 UCI benchmark datasets and one real world dataset. Particularly, in the case of the real medical diagnosis, the cost values are set according to the experts suggestions and practical medical prices.

5.1. An illustrative example

One of UCI dataset Wine is used as an example to illustrate the effect of FCS-RF. First, the feature rank acquired from the RF is used as a baseline, shown in the first row in Table 2. In this case, all features are assumed to share the same feature cost value of 1. From the rank we know the important features (i.e., the top ranked features are more informative). We choose 3 top ranked features (i.e., the 10th, the 7th, and the 13th features) and assign larger costs to them to observe how the cost change affects the rank. Applying FCS-RF, as their cost increases, these three features should have smaller probabilities of being selected and, consequently, their importance and ranking should decline. As shown in Table 2, the new feature ranks produced by FCS-RF demonstrate that the original top 3 ranked features (highlighted in bold) move backward in the rank, as we increase their costs,

When their cost reaches 20 or more, the probability of the top 3 features being selected becomes so small that the original most important three features are listed at the bottom in the rank and become the least important features. It should be noted that the position of the features in the rank are not fixed even for the same given parameters. As each time the forest is constructed, there is some randomness in the process and the calculated importance score for the features may have slight differences, affecting the final rank.

5.2. Experiments on UCI datasets

5.2.1. Dataset description

Table 3 shows the summary of the datasets used in the experiments. The last four datasets are characterized with high dimension and small size. DLBCL and Leukemia dataset have more than 10 thousands features. To illustrate the cost-based feature selection process, we performed a pre-processing on these two datasets and remain part of original features.

5.2.2. Experimental setup

Currently there is a shortage of benchmark datasets with true and accurate feature costs, but it is possible to artificially generate the cost

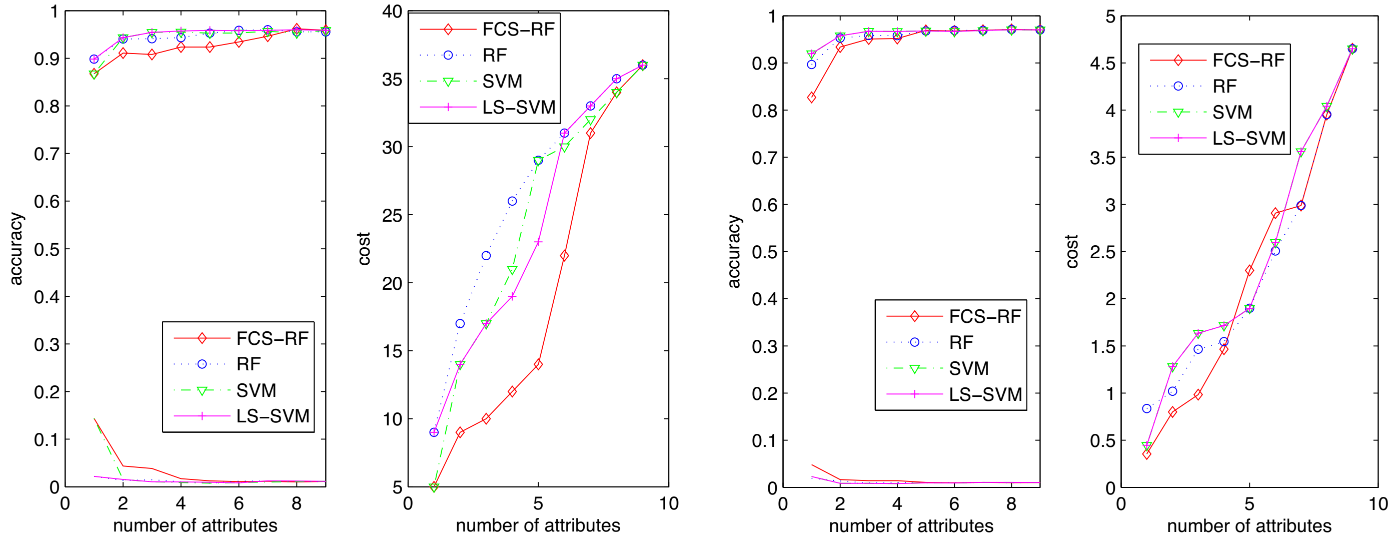
Table 3
The UCI datasets used in the experiments.

Dataset	Features	Size	Classes
Wine	12	178	3
Cancer	9	699	2
Heart	12	270	2
House	16	435	2
Ionosphere	34	351	2
Sonar	60	208	2
DLBCL	114	77	2
Leukemia	136	72	2
Colon	2000	62	2
SRBCT	2308	63	4

through some cost-setting strategies to simulate the true situations. In particular, we set the feature cost in the following two ways. One way is to set the cost randomly: a random number between 0 and 1 is assigned for all features.

$$c_i = random(0,1), i = 1, 2, \ldots, n. \tag{12}$$

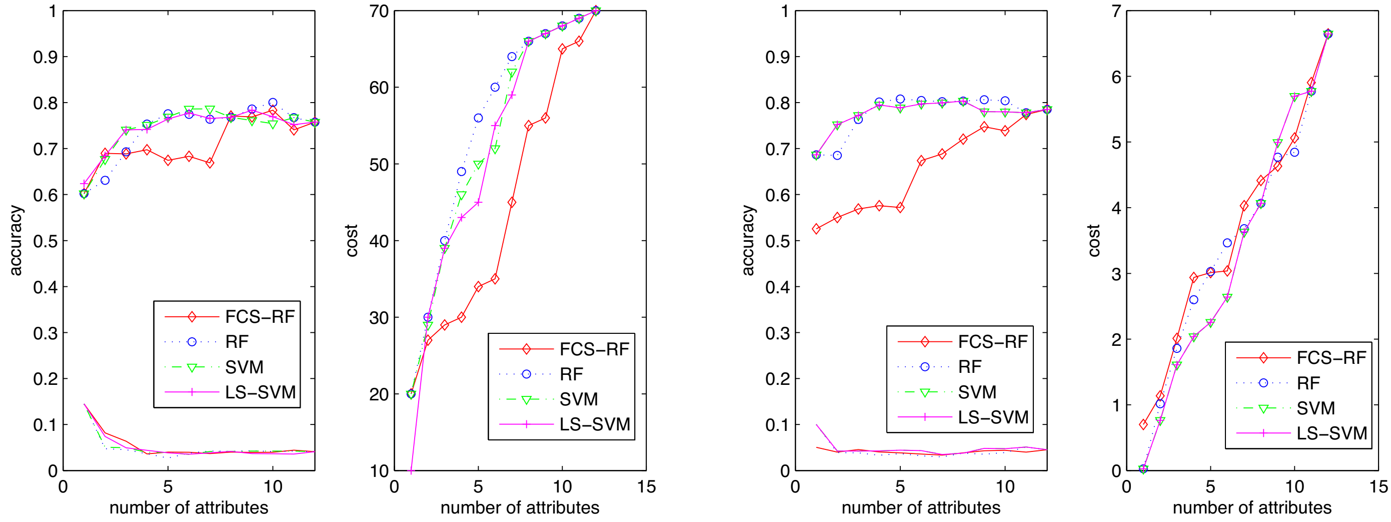
This setting is to simulate the situations when we are entirely ignorant of the specific values of the feature cost. The other way is to generate the feature cost based on the importance score produced by RF as shown in Eq.(11). This setting especially complies with our intuition that more important things tend to be more expensive. It is necessary to point out that the specific value of each feature cost is not important. From the FCS-RF algorithm, the relative magnitude between the cost values determines the possibility distribution and affects the selection process. We conduct a comparative experiment between the proposed FCS-RF and three feature selection methods, SVM-RFE, LS-SVM (Least Squares SVM) and the ordinary random forest (they are very common and typical methods, and have been applied in many fields. For simplicity, we denote these three compared methods as SVM, LS-SVM and RF in the experiments. SVM(-RFE), LS-SVM and RF are very common and typical methods, and they have



(a) cost-setting based on feature importance

(b) cost-setting based on random

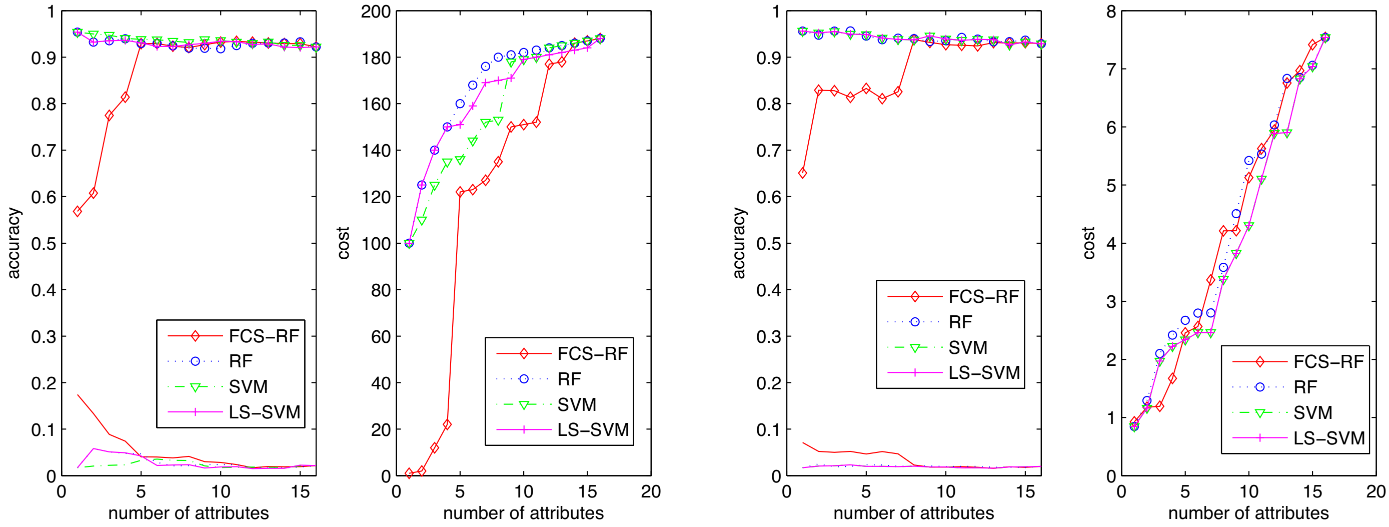
Fig. 1. Results on dataset Cancer.



(a) cost-setting based on feature importance

(b) cost-setting based on random

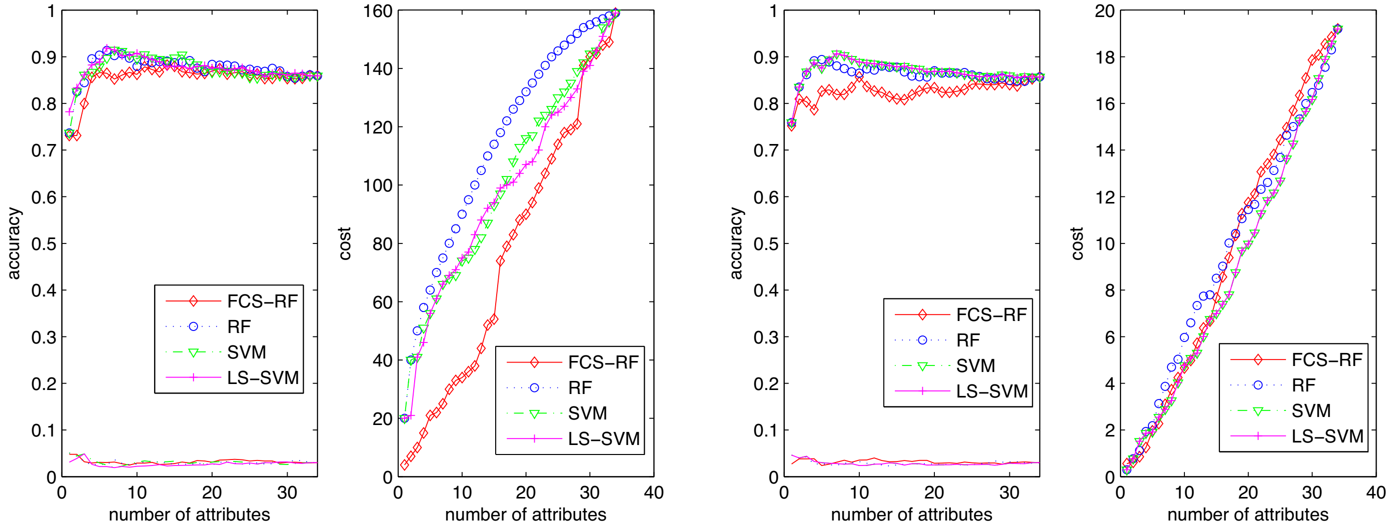
Fig. 2. Results on dataset Heart.



(a) cost-setting based on feature importance

(b) cost-setting based on random

Fig. 3. Results on dataset House.



(a) cost-setting based on feature importance

(b) cost-setting based on random

Fig. 4. Results on dataset Ionosphere.

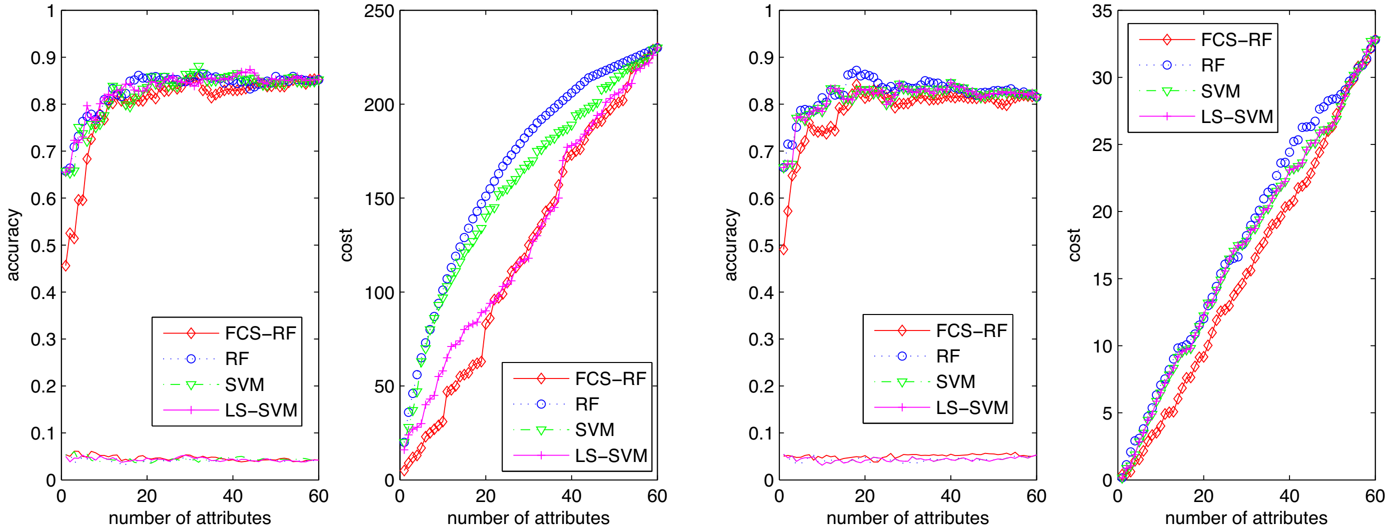
applied in many fields). Given the same feature cost, four different feature ranks could be obtained through FCS-RF, RF, SVM, and LS-SVM respectively.

For comparison, we select the same amount of the top ranked features in each rank simultaneously and use the selected features to construct the KNN classifiers. The accuracy of KNN can measure the quality (or distinguishing ability) of the selected features and is regarded as one key criterion of the features. We will simply call the accuracy of classifier as the corresponding accuracy of features. As shown in Fig. 1–9 the accuracy and cost of the selected features are separately computed and plotted as the number of the used features grows from 1 to all. Because the cost values assigned to the features are different, the scales of the cost axis may be different in the figures. The standard deviation of the corresponding accuracy is also plotted at the bottom of the chart of accuracy. In the experiments, the KNN's accuracy is averaged over 30 rounds and in each round, 70% randomly selected samples are as training set and others are as test set. The values for each feature, in the KNN construction, would be first normalized to the range [0 1] in order to guarantee the effect of KNN.

5.2.3. Results analysis

From Fig. 1–9, we can summarize some common observations as follows.

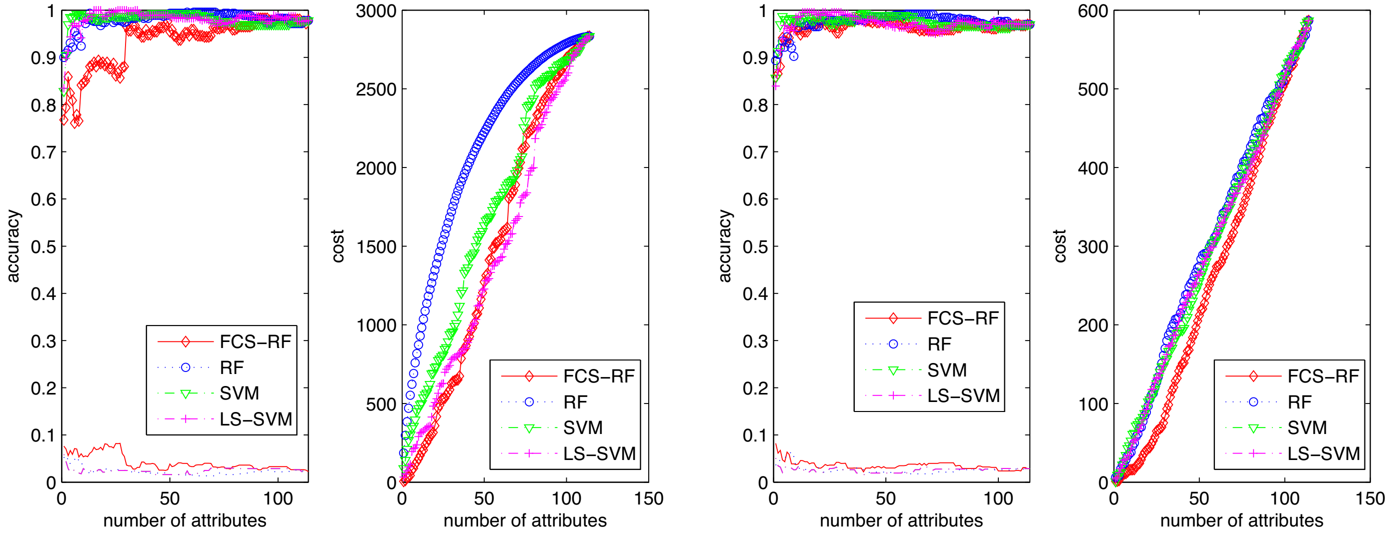
(1) The accuracy-curves of FCS-RF, RF, SVM, and LS-SVM have similar trend in both cases of the cost-setting strategy. Starting with a growing period and then reaching a plateau, the trend complies with the effect of feature selection. On almost all cases, it only needs a very small portion of the full feature set to reach a satisfactory accuracy (stable stage). The FCS-RF curve may lower than the others when the number of the used features is not large enough, seen in Figs. 2(a),3(a),5(a), and6(a), and this is owing to the effect of feature cost influencing the selection of top features. Note that the variation is not bound to happen, e.g., in Figs. 1 and 4, because of two conditions: some features are vital to the classification and the cost distribution hardly exerts any influence to their ranking (unless the cost difference approximates to infinity); some suboptimal but cheaper features could replace the optimal features and maintain the accuracy. As more features are added, the originally lower FCS-RF curve gradually reaches the same height as the others. This is because the



(a) cost-setting based on feature importance

(b) cost-setting based on random

Fig. 5. Results on dataset Sonar.



(a) cost-setting based on feature importance

(b) cost-setting based on random

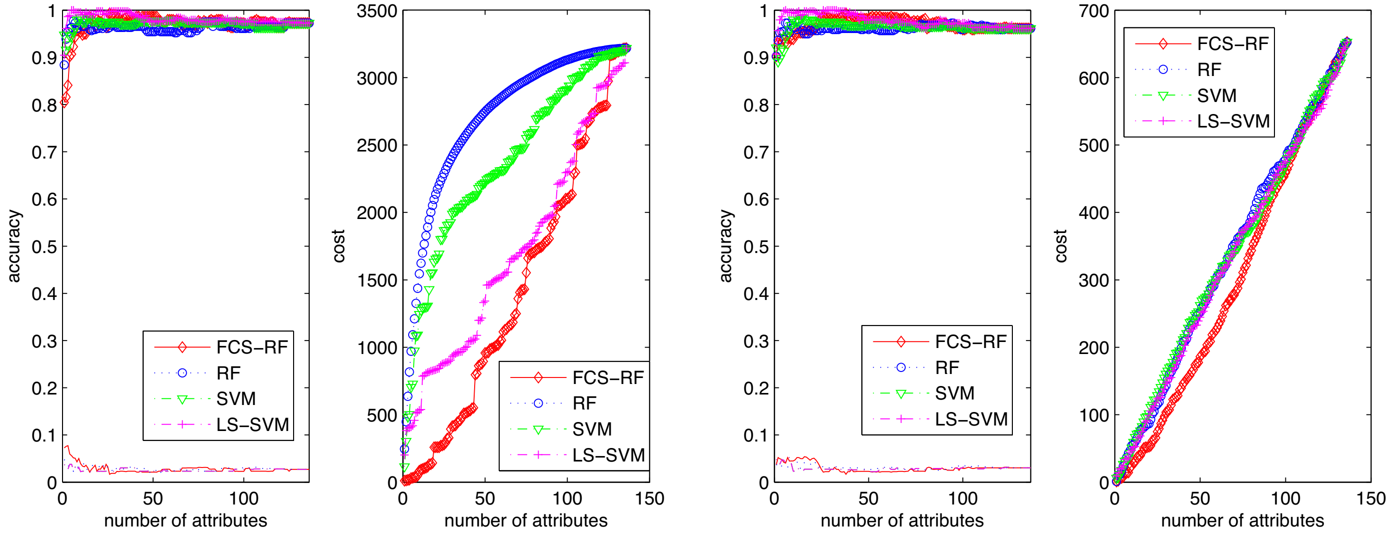
Fig. 6. Results on dataset DLBCL.

overlapping rate of the three feature ranks increases. The decline of the accuracy-curves on Colon and SRBCT indicates that some features may bring a negative influence on classification.

(2) In terms of the cost, on most datasets the FCS-RF curve is lower than the others, especially in the early stage of the horizontal axis. This shows that FCS-RF tends to select low-cost features. For the feature subsets of the same size produced by FCS-RF, RF, SVM, and LS-SVM, the FCS-RF subset usually has smaller cost, and at the same time, its accuracy is similar or only slightly inferior to its comparisons (the exceptions are Figs. 3(a) and 6(a)). This means the FCS-RF achieves finding the both informative and low-cost feature subsets. For the cases in Figs. 3(a) and 6(a), in order to guarantee a good accuracy, the size of the feature subset from FCS-RF has to be larger than the subsets from RF, SVM and LS-SVM. In spite of this, the FCS-RF subset may still be cheaper, e.g., in the Fig. 3(a), the FCS-RF can reach the accuracy of 91% using the top 11 features, but the total cost of these features is even less than the cost of the first feature in the other two ranks (see the right-sided plot of Fig. 3(a)). Overall, the three cost-curves get increasingly close until converge when all features are used. This phenomenon can be explained by the fact that adding more features can increase the overlapping rate of the three feature subsets. When all features are used, the total cost, obviously, is a constant.

(3) The standard deviation of the accuracy can be excessive when there are only very few features, as seen prominently in Figs. 1–3, or on the high-dimension small size datasets. Considering only a couple of features are used, the classifier can become very sensitive to the data perturbation. The date sets Colon and SRBCT have a high standard deviation due to their data characteristics. However, it is still possible to stabilize the accuracy, as shown in Fig. 9, by choosing the proper features.

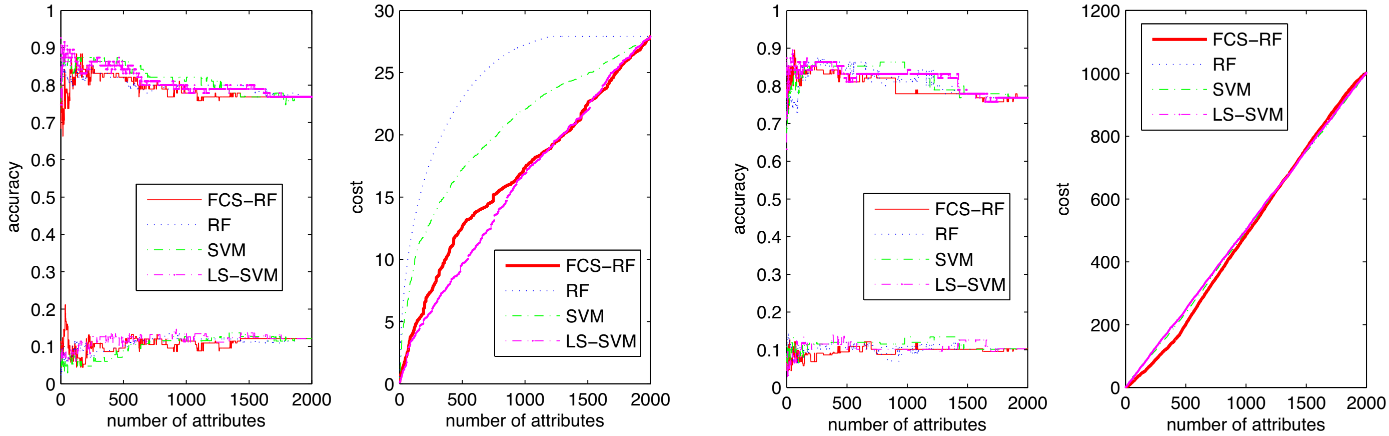
(4) There are generally three ways to solve SVM optimization models, including QP (quadratic programming), SMO (Sequential Minimal Optimization method), and LS (least-squares method). Ref.



(a) cost-setting based on feature importance

(b) cost-setting based on random

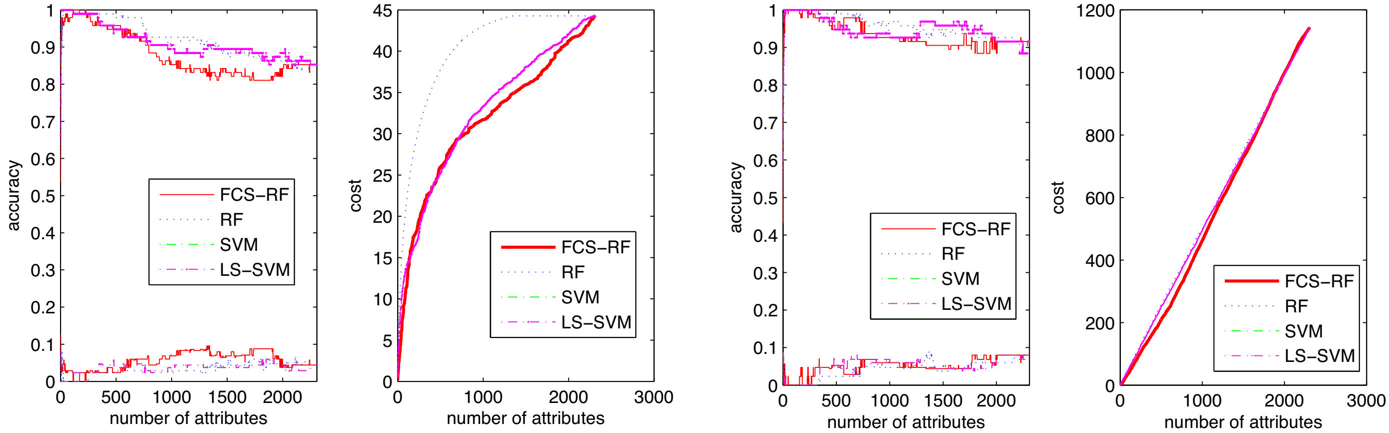
Fig. 7. Results on dataset Leukemia.



(a) cost-setting based on feature importance

(b) cost-setting based on random

Fig. 8. Results on dataset Colon.



(a) cost-setting based on feature importance

(b) cost-setting based on random

Fig. 9. Results on dataset SRBCT.

Table 4
Comparison when setting cost based on importance.

Dataset	Approach	Features	Accuracy%	Cost
Cancer	No FS	9	97	100
	FCS-RF	2	94	33
	RF	2	95	47
	SVM	2	95	39
Heart	No FS	12	78	100
	FCS-RF	6	80	49
	RF	4	80	70
	SVM	4	80	70
House	No FS	16	93	100
	FCS-RF	12	94	98
	RF	2	95	66
	SVM	2	95	61
Ionosphere	No FS	34	84	100
	FCS-RF	2	86	3
	RF	2	84	25
	SVM	2	84	25
Sonar	No FS	60	82	100
	FCS-RF	12	80	19
	RF	7	82	34
	SVM	7	80	31
DLBCL	No FS	114	97	100
	FCS-RF	30	96	22
	RF	10	97	31
	SVM	8	98	14
Leukemia	No FS	136	97	100
	FCS-RF	26	99	9
	RF	6	98	34
	SVM	8	98	34
Colon	No FS	2000	74 ± 11	100
	FCS-RF	104	87 ± 7	14
	RF	5	86 ± 10	8
	SVM	29	88 ± 8	16
SRBCT	No FS	2308	78 ± 9	100
	FCS-RF	38	100	11
	RF	31	100	26
	SVM	12	100	9

Table 5
Comparison when setting cost based on random.

Dataset	Approach	Features	Accuracy%	Cost
Cancer	No FS	9	96	100
	FCS-RF	3	95	24
	RF	3	95	32
	SVM	2	95	28
Heart	No FS	12	77	100
	FCS-RF	5	79	29
	RF	4	79	39
	SVM	4	80	31
House	No FS	16	93	100
	FCS-RF	2	95	11
	RF	2	95	17
	SVM	2	95	15
Ionosphere	No FS	34	84	100
	FCS-RF	4	90	6
	RF	4	90	10
	SVM	4	90	10
Sonar	No FS	60	83	100
	FCS-RF	16	86	15
	RF	16	86	31
	SVM	12	84	34
DLBCL	No FS	114	98	100
	FCS-RF	19	98	7
	RF	10	98	6
	SVM	9	98	8
Leukemia	No FS	136	97	100
	FCS-RF	25	99	11
	RF	13	97	10
	SVM	15	99	12
Colon	No FS	2000	72 ± 7	100
	FCS-RF	76	89 ± 6	2
	RF	5	88 ± 7	0
	SVM	31	89 ± 8	2
SRBCT	No FS	2308	76 ± 11	100
	FCS-RF	30	100	1
	RF	30	100	2
	SVM	12	100	1

[30] studies the relationship between SVM and LS-SVM, and the analysis showed that under some mild conditions, LS-SVM is equivalent to SVM. Our experiment results are consistent with their analysis. From Fig. 1–9 we can see that, SVM and LS-SVM show similar performance on most of the data sets. Especially under the condition of random cost-setting, LS-SVM and SVM almost have the same predict accuracy and feature costs. Therefore, in our following analysis, we only give the performance comparison results of FCS-RF, RF, and SVM.

To illustrate the effectiveness of the proposed method from an intuitive perspective, we draw the tables to compare the cost and the accuracy of the feature subset produced by FCS-RF, RF and SVM, as shown in Tables 4 to 5. According to the Figs. 1–9, for each dataset, certain optimal feature subset could be determined. For instance, in Fig. 1(a), for the dataset Cancer, the top two features in the rank, could be used as a subset replacing the set of all features, as the two selected features are able to guarantee a reasonable accuracy and standard deviation. By contrast, a single feature will cause a high standard deviation, although the related average accuracy seems promising.

The size of the optimal subset for different methods can be different. For the dataset House in the Fig. 3(a), the optimal subset generated through FCS-RF is nearly 6 times larger than the other two subsets. The total cost and the original accuracy of all features (No FS) are also presented in the tables. For readability, the standard deviation is appended only if it is over 5%. The cost has been normalized by multiplying some value so as to set the cost of all features 100. From the tables, the selected feature subset after is usually much smaller than the complete set, but it could be used to construct a classifier with equal or higher accuracy. In the column of Cost, the cost consumed in FCS-RF is the lowest on most occasions. The advantage may be insignificant on DLBCL or SRBCT. For Colon, the accuracy becomes so unstable that the comparison of costs is of little meaning. On the dataset House in Table 4, the cost of FCS-RF is 98, much higher than that of the compared methods. It could be explained that the 12th features in the FCS-RF rank is extremely expensive, as shown in Fig. 3(a). Removing it from the subset will make the cost drop from 98 to 45, but also reduce the accuracy from 94% to 91%. Therefore, this clearly shows the tradeoff between the cost and the accuracy.

In Table 4, the "accuracy" means the KNN classifiers accuracy on the test set (e.g. accounts for about 30% of all instances in a data set). We note that there is an anomalous entry 0 in the case of Colon. The reason is that each feature cost value is a random between 0 and 1, and the cost sum of the top 5 features selected by RF happens to be under 0.5. We should stress the point that the importance-based cost strategy, in effective, reflects the worst feature cost distribution (the best distribution is that the feature cost is inversely proportional to the features importance or discriminant ability). As a result, compared by the random costs, the performance of FCS-RF using the importance-based cost distribution should be less advantageous than its comparisons. This assumption is confirmed by the figures and tables: in the random cost-setting mode, the advantage of FCS-RF is more significant.

5.3. Experiment on real world dataset

In medical diagnosis, the costs of measuring the basic characteristics (e.g. age, gender, and weight) of a patient are cheap, while the costs of obtaining some other characteristics (e.g. blood tests, genotype evaluations) are expensive. The difference of these costs may be quite large. In this case, both the physicians and patients need to evaluate the tradeoff between the potential benefits and costs of different

Table 6
The details of the dataset Hepatitis.

No.	Feature	Description	Cost
1	Gender	The gender of the patient	1
2	Age[a]	The age of the patient	1
3	G[b]	Intrahepatic inflammatory activity	50
4	S[b]	Liver fibrosis	50
5	ALT[c]	Alanine aminotransferase	10
6	AST[c]	Aspartate aminotransferase	10
7	HBV-DNA	The maximum of HBV-DNA	20
8	Gene Type[b]	The gene type of HBV	40
9	MTCT	Whether mother-to-child transmission	5
10	IFN Type	The type of the interferon	1
11	IFN Dose	The dose of the interferon	1

[a] Less than 15 denoted as 1, between 15 and 24 as 2, between 25 and 44 as 3, and over 44 as 4

[b] Divided into 4 grades or categories

[c] scriptsize Less than 41 U/L denoted as 0, between 41 and 80 U/L as 1, between 81 and 200 U/L as 2, between 201 and 400 as 3, and over 400 as 4

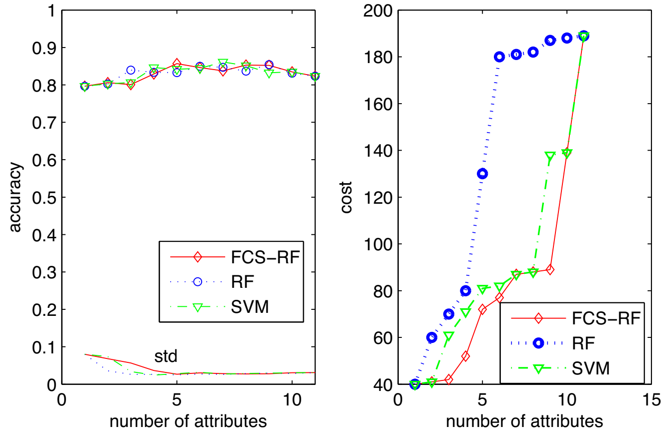


Fig. 10. Results on the dataset Hepatitis.

treatment options. In this section, we apply the proposed method to interferon-α (IFN-α) treatment of chronic hepatitis B (CHB) to help find the significant but lower-cost predictive factors.

5.3.1. Problem description

Chronic hepatitis B (CHB) is one of the most refractory diseases and IFN-α treatment is one of available antiviral options. Since the response to IFN-α therapy varies from person to person (i.e., the therapy is effective to someone but ineffective to others), response prediction for individual patients is thus important. That is, finding the key response factors from the possible candidates (e.g., the basic characteristics and other medical test results patients) can meanwhile lower the possible testing cost.

We apply our proposed method to this problem and verify its effectiveness on a real world dataset, which includes a total of 382 patients treated by IFN-α. The features in this set as well as their descriptions are shown in Table 6.

The class label belongs to $\{1, -1\}$, where 1 denotes the IFN-α improves the patients condition significantly and -1 denotes the IFN-α doesn't work. There are 267 samples labeled 1 and 115 samples labeled -1. The feature cost is estimated by the experts according to a variety of factors such as price, time, possible side effect, and operability. To quantify the cost values, an integer between 1 and 50 is assigned to each feature, as shown in the last column in Table 6.

5.3.2. Experiments and analysis

In the experiments, the dataset is first normalized and then we use FCS-RF, RF, and SVM to perform feature selection and produce three feature ranks, respectively, as shown in Table 7. We use KNN to estimate the quality of the features and compute their cost. The results are shown in Fig. 10.

From Table 7, for all three approaches, the 8th feature, Gene Type, is the most important feature. However, using a single feature may result in large standard deviation, as shown in Fig. 10, and we should combine more features to improve the stability. Adding more features will improve the accuracy slightly, as shown in Fig. 10. But with regard to the cost, when the number of the used features is fixed, the FCS-RF curve is lower than the other two curves. In other words, FCS-RF selects the cheaper features in contrast to RF and SVM. This shows that FCS-RF tends to select the lower-cost features among those of similar quality. Considering both the accuracy and its stability (the standard deviation), the minimal feature subset, generated by FCS-RF, RF, and SVM, respectively, are {#8, #2, #11}, {#8, #7}, and {#8, #10, #7}, with the corresponding total cost of 42, 60, 61. Our proposed FCS-RF outperforms the other two approaches.

Table 7
Three feature ranks on Hepatitis.

FS approach	Feature rank										
FCS-RF	8	2	11	6	7	9	5	10	1	4	3
RF	8	7	6	5	4	3	2	10	9	1	11
SVM	8	10	7	6	5	1	9	11	4	2	3

6. Conclusions

In this paper, we propose a feature-cost-sensitive random forest to perform feature selection. Different from traditional feature selection methods, the feature cost is taken into account and the feature subset produced through FCS-RF can be both low-cost and informative. Specifically, we convert the feature cost into a probability vector and incorporate it into the tree growth. By means of ensemble, the proposed method is able to reduce the cost while preserving the accuracy. It should be emphasized that only if the accuracy is acceptable, the decrease of cost is meaningful. As a result, our proposed method are especially useful in the cases where there are some redundant and expensive features in the original data where our algorithm could conduct effective cost-sensitive feature selection. The experiments results verified the effectiveness of our proposed method and demonstrated that the combination of the probability section and the ensemble is effective.

As future work, we will explore more advanced methods to consider the feature costs. One direction is transforming feature costs to constraints and embedding them into the process of feature selection to solve a global optimization problem. Another interesting direction is taking into account various costs (e.g. feature cost, misclassification cost et.al.) simultaneously, and a challenge here is how to make a tradeoff among various costs in practical problems.

Acknowledgment

This work is supported by Natural Science Foundation of China under grant no. 61503313, and the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology), grant no. 30920140122007.

References

[1] P.N. Tan, M. Steinbach, V. Kumar, Introduction to data mining, Second Edition, Posts & Telecom Press, 2006.

[2] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[3] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression[J], Bioinformatics 20 (15) (2004) 2429–2437.

[4] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance,and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.

[5] Q. Zhou, H. Zhou, Q. Zhou, et al., Structure damage detection based on random forest recursive feature elimination[J], Mech Syst Signal Process 46 (1) (2014) 82–90.

[6] P. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.

[7] P. Langley, Selection of relevant features in machine learning, AAAI Fall Symp. Relev. (1994) 140–144.

[8] I. Guyon, J. Weston, S. Barnhill, Gene Selection for Cancer Classification using Support Vector Machines, Mach. Learn. 46 (2002) 389C422.

[9] C. Elkan, The foundations of cost-sensitive learning, Proceedings of the International joint conference on artificial intelligence, 17(1), LAWRENCE ERLBAUM ASSOCIATES LTD, 2001, pp. 973–978.

[10] P. Domingos, Metacost: A general method for making classifiers cost-sensitive, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1999, pp. 155–164.

[11] P.N. Turney, Types of cost in inductive concept learning, Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learing, Stanford University, California, 2000.

[12] P. Turney, Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm, J. Artif. Intell. Res. (JAIR) (1995) 2.

[13] F. Min, Q. Liu, A hierarchical model for test-cost-sensitive decision systems, Inf. Sci. 179 (14) (2009) 2442–2452.

[14] C.X. Ling, Q. Yang, J. Wang, et al., Decision trees with minimal costs, Proceedings of the twenty-first international conference on Machine learning,, ACM, 2004, p. 69.

[15] S. Sheng, C.X. Ling, Hybrid cost-sensitive decision tree, Knowledge Discovery in Databases: PKDD 2005,, Springer Berlin Heidelberg, 2005, pp. 274–284.

[16] C.X. Ling, V.S. Sheng, Q. Yang, Test strategies for cost-sensitive decision trees,, IEEE Trans. Knowl. Data Eng. 18 (8) (2006) 1055–1067.

[17] H. He, H. Daumé III, J. Eisner, Cost-sensitive dynamic feature selection, ICML Workshop on Inferning: Interactions between Inference and Learning, Edinburgh., 2012.

[18] S. Ji, L. Carin, Cost-sensitive feature acquisition and classification, Patt. Recognit. 40 (5) (2007) 1474–1485.

[19] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (5) (1999) 988–999.

[20] C. Orsenigo, C. Vercellis, Multivariate classification trees based on minimum features discrete support vector machines, IMA J. Manag. Math. 14 (3) (2003) 221–234.

[21] M. Núnez, The use of background knowledge in decision tree induction, Mach. Learn. 6 (3) (1991) 231–250.

[22] M. Tan, J.C. Schlimmer, Cost-sensitive concept learning of sensor use in approach and recognition, Proceedings of the sixth international workshop on Machine learning, Morgan Kaufmann Publishers Inc., 1989, pp. 392–395.

[23] S.W. Norton, Generating better decision trees, in: Proceedings of the IJCAI, 89, 1989, pp. 800–805.

[24] R. Genuer, J.M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognit. Lett. 31 (14) (2010) 2225–2236.

[25] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[26] H.R. Zhang, F. Min, Three-way recommender systems based on random forests, Knowl. Based Syst. (2015).

[27] Q. Wu, Y. Ye, H. Zhang, et al., ForesTexter: an efficient random forest algorithm for imbalanced text categorization, Knowl. Based Syst. 67 (2014) 105–116.

[28] C.C. Yeh, F. Lin, C.Y. Hsu, A hybrid KMV model, random forests and rough set theory approach for credit rating, Knowl. Based Syst. 33 (2012) 166–172.

[29] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[30] J. Ye, T. Xiong, SVM versus least squares SVM, in: International Conference on Artificial Intelligence and Statistics, 2007, pp. 644–651.

[31] Y. Zhang, C. Ding, T. Li, Gene selection algorithm by combining reliefF and mRMR, BMC Genom. 9 (Suppl 2) (2008) S27.

[32] S. Zhu, D. Wang, K. Yu, T. Li, Y. Gong, Feature Selection for gene expression using model-based entropy, IEEE/ACM Trans. Comput. Biol. Bioinformatics 7 (1) (2010) 25–36.