ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20220823104912
Original scientific paper

Research on Feature Selection Methods based on Random Forest

Zhuo WANG

Abstract: Aiming to deal with the irrelevant or redundant features, this paper proposes eight kinds of feature selection methods. The first seven feature selection methods include CART and Random Forests (CART-RF), CHIAD and Random Forests (CHIAD-RF), SVM and Random Forests (SVM-RF), Bayesian Network and Random Forests (BN-RF), neural Network and Random Forests (NN-RF), K-Means and Random Forests (K-Means-RF) and Kohonen and Random Forests (Kohonen-RF). These methods use CART, CHAID, SVM, BN, NN, K-Means and Kohonen to evaluate the importance and ranking of features, and then obtain feature subsets through RF algorithm. The eighth method is named hybrid integration methods and random forests (Integrate-RF). Integrate-RF uses the average importance of the seven methods and the optimal features subset can be selected based on the OOB data classification error rate. Experimental results indicate that feature selection methods proposed in this article can effectively select features and reduce the data dimension.

Keywords: feature selection; irrelevant; random forest; redundant

1 INTRODUCTION

With the data dimension increases, there are more and more irrelevant features and redundant features; thereby the learner cannot fully obtain the effective information from data, reducing the performance of the learner. Especially for small-sample and high-dimensional data, it is easy for "dimensional disaster" and over-fitting phenomenon to occur. To deal with such a problem, feature selection [1-2] is an effective method; it removes irrelevant features and redundant features to reduce the number of features, and finds optimal feature subset which is the smallest but can improve the learner's recognition to the maximum. Feature selection [3] is very useful to solve dimension reduction of high dimensional data, especially for the modelling of small-sample and high-dimensional data.

Feature selection [4] is a process of cyclically searching for the optimal feature subset; it mainly includes feature importance evaluation, feature subset generation, feature subsets evaluation, search termination condition, results verification and confirmation. Among them, feature importance evaluation [5-11] is one of the most important steps to feature selection. Feature selection algorithms can be divided into supervised feature selection [12] and unsupervised feature selection [13], and can also be divided into four categories [14-16]: filter methods, wrapper methods, embedded methods and hybrid methods.

The paper proposes eight kinds of feature selection methods, the eighth feature selection method using hybrid integration of difference models and random forests (Integrate-RF). The difference models include CART [17], CHAID [18], SVM [19], Bayesian Networks (BN) [20], Neural Networks (NN) [21], K-Means [22] and Kohonen [23]. Integrate-RF [24] method obtains ordering of feature importance by integrating various supervised and unsupervised feature evaluation methods, and generates feature subset by forward search strategy, then evaluates the feature subsets in terms of the minimum OOB error of the random forests, feature number of the lowest OOB error, the average OOB error, the variance of OOB error rate.

2 METHOD

2.1 Feature-Importance-Evaluation Methods

Hybrid feature selection method includes feature importance evaluation, feature subsets generation and feature subsets evaluation. Feature importance evaluation is one of the most important steps to feature selection. Integrate-RF method obtains ordering of feature importance by hybrid integration of difference models. CART, CHAID, SVM, Bayesian Networks (BN), and Neural Networks (NN) belong to supervised embedded feature selection algorithms. K-Means and Kohonen are unsupervised embedded feature selection algorithms, they are all available for feature-importance-evaluation.

2.1.1 Supervised Methods

Feature importance can be determined by computing the reduction in variance of the target attributable to each feature, via a sensitivity analysis. This method of computing feature importance is used in the following models: CART, CHAID, SVM, BN, NN, when Y is the target, X_j is feature, where $j = 1, \ldots, k$, k is number of features, $Y = f(X_1, X_2, \ldots, X_k)$. Model for Y is based on features X_1 through X_k. Features are ranked according to the sensitivity measure defined as follows.

$$S_i = \frac{V_i}{V(Y)} = \frac{V\left(E\left(Y|X_i\right)\right)}{V(Y)} \tag{1}$$

where $V(Y)$ is the unconditional output variance. In the numerator, the expectation operator E calls for an integral over X_{-i}; that is, over all factors but X_i, then the variance operator V implies a further integral over X_i. Feature importance is then computed as the normalized sensitivity.

$$VI_i = \frac{S_i}{\sum_{j=1}^{k} S_j} \tag{2}$$

Classification and regression tree (CART) splits the data into two subsets so that the samples within each subset are more homogeneous than in the previous subset. It is a

recursive process; each of those two subsets is then split again, and the process is repeated until the homogeneity criterion is reached or some other stopping criterion is satisfied. The same predictor field may be used several times at different levels in the tree. It uses surrogate splitting to make the best use of data with missing values. CART is quite flexible. It allows unequal misclassification costs to be considered in the tree growing process. It also allows you to specify the prior probability distribution in a classification problem.

CHAID stands for Chi-squared Automatic Interaction Detector. It is a highly efficient statistical technique for segmentation, or tree growing. Using the significance of a statistical test as a criterion, CHAID evaluates all of the values of a potential predictor field. It merges values that are judged to be statistically homogeneous (similar) with respect to the target variable and maintains all other values that are heterogeneous (dissimilar). It then selects the best predictor to form the first branch in the decision tree, such that each child node is made of a group of homogeneous values of the selected field. This process continues recursively until the tree is fully grown. The statistical test used depends upon the measurement level of the target field. If the target field is continuous, an *F* test is used. If the target field is categorical, a chi-squared test is used. CHAID is not a binary tree method; that is, it can produce more than two categories at any particular level in the tree. Therefore, it tends to create a wider tree than the binary growing methods. It works for all types of variables, and it accepts both case weights and frequency variables. It handles missing values by treating them all as a single valid category.

The Support Vector Machine (SVM) is a supervised learning method that generates input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function or a regression function. For classification, nonlinear kernel functions are often used to transformed input data to a high-dimensional feature space in which the input data become more separable compared to the original input space. Maximum-margin hyperplanes are then created. The produced model depends on only a subset of the training data near the class boundaries. Similarly, the model produced by Support Vector Regression ignores any training data that is sufficiently close to the model prediction. (Support Vectors can appear only on the error tube boundary or outside the tube.)

Bayesian Networks (BN) provides a succinct way of describing the joint probability distribution for a given set of random variables. Let V be a set of categorical random variables and $G = (V, E)$ be a directed acyclic graph with nodes V and a set of directed edges E. A Bayesian network model consists of the graph G together with a conditional probability table for each node given values of its parent nodes. Given the value of its parents, each node is assumed to be independent of all the nodes that are not its descendents. The joint probability distribution for variables V can then be computed as a product of conditional probabilities for all nodes, given the values of each node's parents. Given set of variables V and a corresponding sample dataset, we are presented with the task of fitting an appropriate Bayesian network model. The task of determining the appropriate edges in the graph G is called structure learning, while the task of estimating the conditional probability tables given parents for each node is called parameter learning.

Neural networks (NN) predict a continuous or categorical target based on one or more predictors by finding unknown and possibly complex patterns in the data. The multilayer perceptron (MLP) is a feed-forward, supervised learning network with up to two hidden layers. The MLP network is a function of one or more predictors that minimizes the prediction error of one or more targets. Predictors and targets can be a mix of categorical and continuous fields.

2.1.2 Unsupervised Methods

This method uses the following models to compute the importance of predictors: K-Means, Kohonen, when Y is target, X_j is predictor, where $j = 1, \ldots, k$, k is number of predictors, $Y = f(X_1, X_2, \ldots, X_k)$, model for Y is based on predictors X_1 through X_k. The importance of feature i is defined as:

$$VI_i = \frac{-\log_{10}(sig_i)}{\max_{j\in\Omega}\left(-\log_{10}(sig_j)\right)} \tag{3}$$

where Ω denotes the set of predictor and evaluation features, sig_i is the significance or p-value computed from applying a certain test, as described below. If sig_i equals zero, set $sig_i = MinDouble$, where $MinDouble$ is the minimal double value. In across clusters, the p-value for categorical feature is based on Pearson's chi-square, the p-value for continuous features is based on an F test. In clusters, the null hypothesis for categorical feature means that the proportion of cases in the categories in cluster j is the same as the overall proportion, the p-value for categorical features is based on Pearson's chi-square. The null hypothesis for continuous features is that the mean in cluster j is the same as the overall mean, the p-value for continuous features is based on Student's t statistic.

The K-Means method is a clustering method, used to group records based on similarity of values for a set of input fields. The basic idea is to try to discover k clusters, such that the records within each cluster are similar to each other and distinct from records in other clusters. K-Means is an iterative algorithm; an initial set of clusters is defined, and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit).

Kohonen models are a special kind of neural network model that performs unsupervised learning. It takes the input vectors and performs a type of spatially organized clustering, or feature mapping, to group similar records together and collapse the input space to a two-dimensional space that approximates the multidimensional proximity relationships between the clusters. The Kohonen network model consists of two layers of neurons or units: an input layer and an output layer. The input layer is fully connected to the output layer, and each connection has an associated weight. Another way to think of the network structure is to think of each output layer unit having an associated center, represented as a vector of inputs to which it most strongly

responds (where each element of the center vector is a weight from the output unit to the corresponding input unit).

2.2 Feature Subset Generation and Evaluation

RF is a classification that comes with a set of CARTs, and introduces random selection features based on Bagging, which can be used for classification, regression and variable importance analysis. It can process data with missing values, and achieves excellent effect for class imbalanced data; besides, it has high stability and strong generalization, and can complete the test internally and get classification errors. Therefore, this paper uses random forests as a classifier to evaluate the feature subset.

3 PREPROCESSING

CART, CHIAD, SVM, BN, NN, K-means, and Kohonen are seven effective machine learning methods. The paper uses these seven methods for feature importance evaluation, random forests evaluates the feature subset, then forms seven feature selection methods; model structure is shown in Fig. 1.

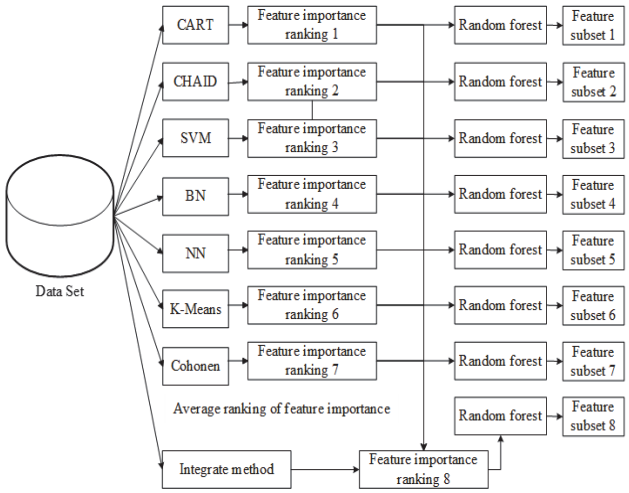


Figure 1 Eight feature selection methods structure

There are seven feature selection methods that are proposed. They include: CART Random Forests (CART-RF), CHIAD Random Forests (CHIAD-RF), SVM Random Forests (SVM-RF), BN Random Forests (BN-RF), NN Random Forests (NN-RF), K-Means Random Forests (K-Means-RF) and Kohonen Random Forests (Kohonen-RF). These methods use CART, CHIAD, SVM, BN, NN, K-means or Kohonen to obtain the feature importance ranking. Then the forward search method is used to generate the feature subset, according to the feature importance. The forward search method starts from the empty subset, and greedily adds the feature with the highest score into the feature subset each time. After adding one feature, the corresponding model is trained and tested; at the end, the classification ability of the feature subset is determined by the OOB error rate of the random forests.

The Integrate-RF model structure is shown in Fig. 1. This method uses seven feature importance evaluation methods to obtain seven feature importance ranking results, takes the average of the seven results as the final feature importance ranking results, and then uses a forward search method to generate a feature subset. Finally, the classification ability of the feature subset is determined by the OOB error rate of the random forests.

4 RESULTS AND DISCUSSION

In order to verify the effectiveness of the eight feature selection methods, the paper designs three sets of experiments. The first set of experiments uses CART, CHAID, SVM, BN, NN, K-Means, and Kohonen to calculate the importance ranking of features, and takes the average of the above seven methods as the feature importance ranking of the Integrate-RF method; the second set of experiments rearranges the features in the data table according to the feature importance to obtain a new data table, and then builds random forests model with decision tree number ranging from 1 - 100 for each new data table, to get the decision tree number when the random forests model is better. The third group of experiments compares the minimum OOB error, OOB error mean and variance, and the number of features in the optimal feature subset for each method with the introduction of features.

4.1 Data Sets Introduction

The experiment uses six sets of UCI classification data sets. A brief description is shown in Tab. 1.

4.2 Feature Importance Ranking

Feature importance evaluation is an important part of the feature selection. This experiment uses seven algorithms of CART, CHAID, SVM, BN, NN, K-Means, and Kohonen in the SPSS Model to obtain seven feature importance ranking results. The average of the feature importance rankings is the feature importance ranking of Integrate-RF. The results are shown in Tab. 2.

4.3 Parameter Selection of Classifier

The number of decision tree has a certain impact on the performance of random forests. The main task in this section is to explore and analyse the relationship between the number of decision tree and the OOB error estimates of random forests, and to find the number of decision tree when the random forest is better. The specific experimental process is as follows: first, reorganize the data table according to the feature importance ranking of each model; second, the number of trees in the random forests starts from 1 to 100, and the step size 20, classify the new data table and get the OOB error estimation rate. The relationship between the OOB error estimation rate of each method and the number of trees for six data sets is shown in Fig. 2.

Table 1 Data sets introduction

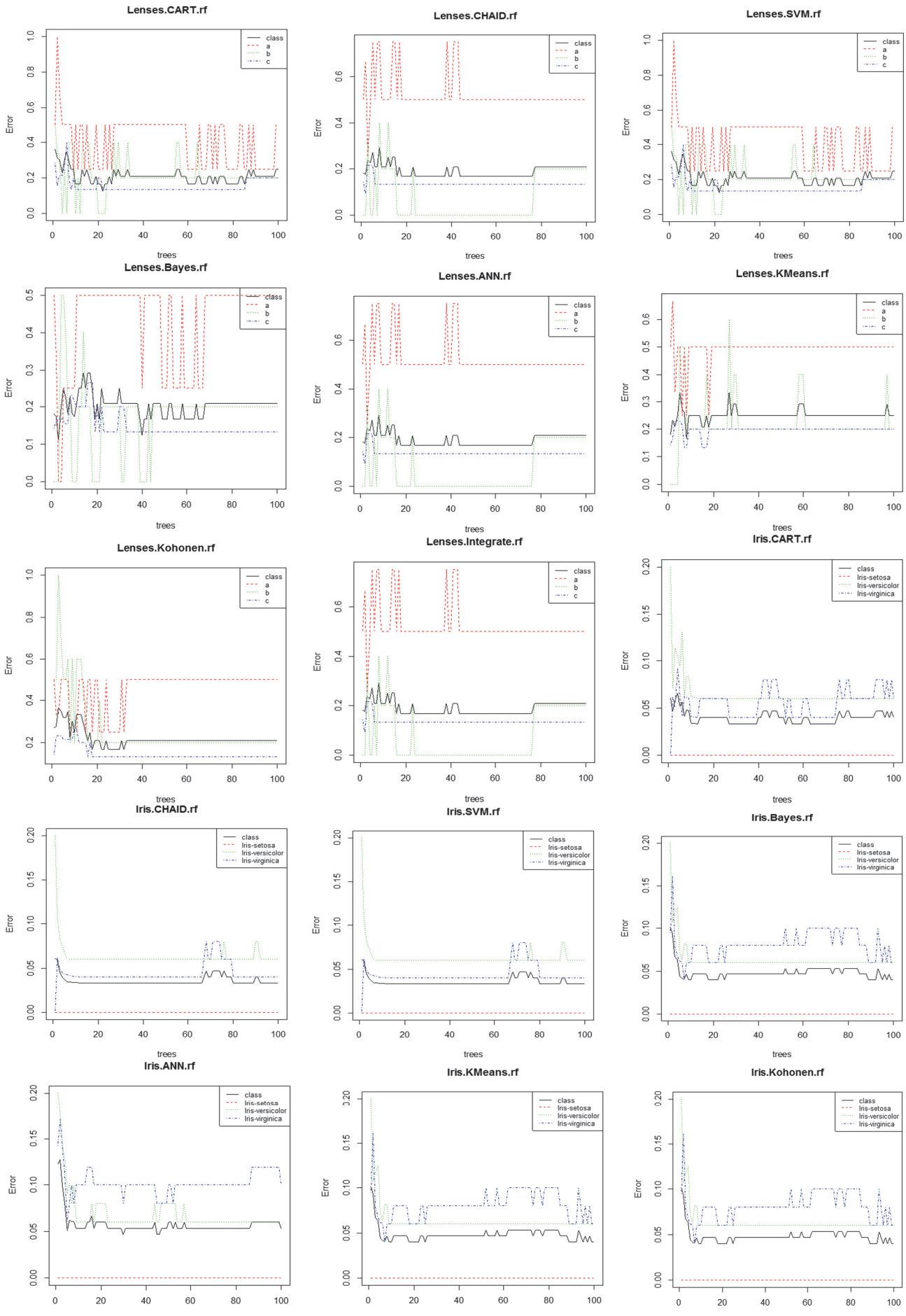
Data Set	Number of instances	Number of features	Number of categories	Missing Values?
Lenses	24	4	3	No
Iris	150	4	3	No
Breast Cancer Wisconsin (Original)	699	9	2	yes
Breast Cancer Wisconsin (Diagnostic)	569	30	2	No
Lung Cancer	32	56	3	yes
SCADI	70	205	7	No

http://archive.ics.uci.edu/ml/datasets.html

Table 2 The results of feature importance ranking

Feature importance ranking（Lenses）								
feature	CART	CHAID	SVM	BN	NN	K-Means	Kohonen	average value sort
astigmatic	2	1	2	3	1	2	2	1
tear	1	2	1	1	2	4	3	2
age	3	3	3	2	3	1	1	3
spectacle	4	3	4	4	4	3	4	4
Feature importance ranking（Iris）								
feature	CART	CHAID	SVM	BN	NN	K-Means	Kohonen	average value sort
sepal width	3	3	3	3	1	4	4	3
petal length	1	2	2	1	2	1	1	1
petal width	2	1	1	2	3	2	2	2
sepal length	4	4	4	4	4	3	3	4
Feature importance ranking (Breast Cancer Wisconsin (Original))								
feature	CART	CHAID	SVM	BN	NN	K-Means	Kohonen	average value sort
Mitoses	9	5	7	6	1	9	9	8
Bare Nuclei	2	2	1	2	2	1	1	1
Bland Chromatin	7	6	4	3	3	4	6	4
Clump Thickness	5	7	2	9	4	8	5	5
Uniformity of Cell Shape	3	9	3	5	5	3	3	3
Uniformity of Cell Size	1	1	9	1	6	2	2	2
Marginal Adhesion	4	3	6	8	7	6	8	7
Normal Nucleoli	6	8	5	4	8	5	4	6
Single Epithelial Cell Size	8	4	8	7	9	7	7	9
Feature importance ranking (Breast Cancer Wisconsin (Diagnostic))								
feature	CART	CHAID	SVM	BN	NN	K-Means	Kohonen	average value sort
feature21	4	9	2	16	4	5	1	1
feature8	13	4	4	1	18	1	6	2
feature23	1	1	3	24	22	4	2	3
feature7	12	9	7	5	14	3	9	4
feature24	5	9	6	25	2	8	7	5
feature28	2	2	1	21	30	2	5	6
feature1	3	9	9	10	28	9	4	7
feature3	8	9	10	30	7	7	3	8
feature14	13	3	17	6	6	15	15	9
feature27	13	9	13	18	5	6	11	10
Feature importance ranking (lung-cancer)								
feature	CART	CHAID	SVM	BN	NN	K-Means	Kohonen	average value sort
a48	17	4	27		3	7	6	1
a54	17	4	9		26	6	3	2
a47	17	4	27		8	7	6	3
a8	17	4	30		21	1	2	4
a19	1	1	6		36	16	18	5
a20	16	4	1		54	3	1	6
a23	10	4	5		29	12	20	7
a6	3	2	16		6	28	27	8
a22	17	4	12		34	18	5	9
a42	17	4	13		12	22	22	10
Feature importance ranking (SCADI)								
feature	CART	CHAID	SVM	BN	NN	K-Means	Kohonen	average value sort
d 53011-2	18	10	3		49	26	18	1
d 5204-1	18	10	22		12	16	49	2
d 5200-0	18	10	20		25	18	51	3
d 5203-1	18	10	21		26	17	50	4
d 5102-4	15	10	56		71	6	4	5
d 5101-4	14	10	56		84	1	1	6
d 5202-1	18	10	56		43	35	10	7
d 5204-2	18	10	56		48	33	8	8
d 571-0	18	10	8		31	25	81	9
d 53011-4	9	10	56		76	12	14	10

Note: Because Breast Cancer Wisconsin (Diagnostic), Lung Cancer, and SCADI have many features, they are only shown the features that the average value sort is top 10.



Note: there are a total of 46 sub-graphs. The first 4 sets of data sets each have 8 sub-graphs, and the last 2 sets of data sets each have 7 sub-graphs.

Figure 2 The relationship between OOB estimate of error rate and the number of trees

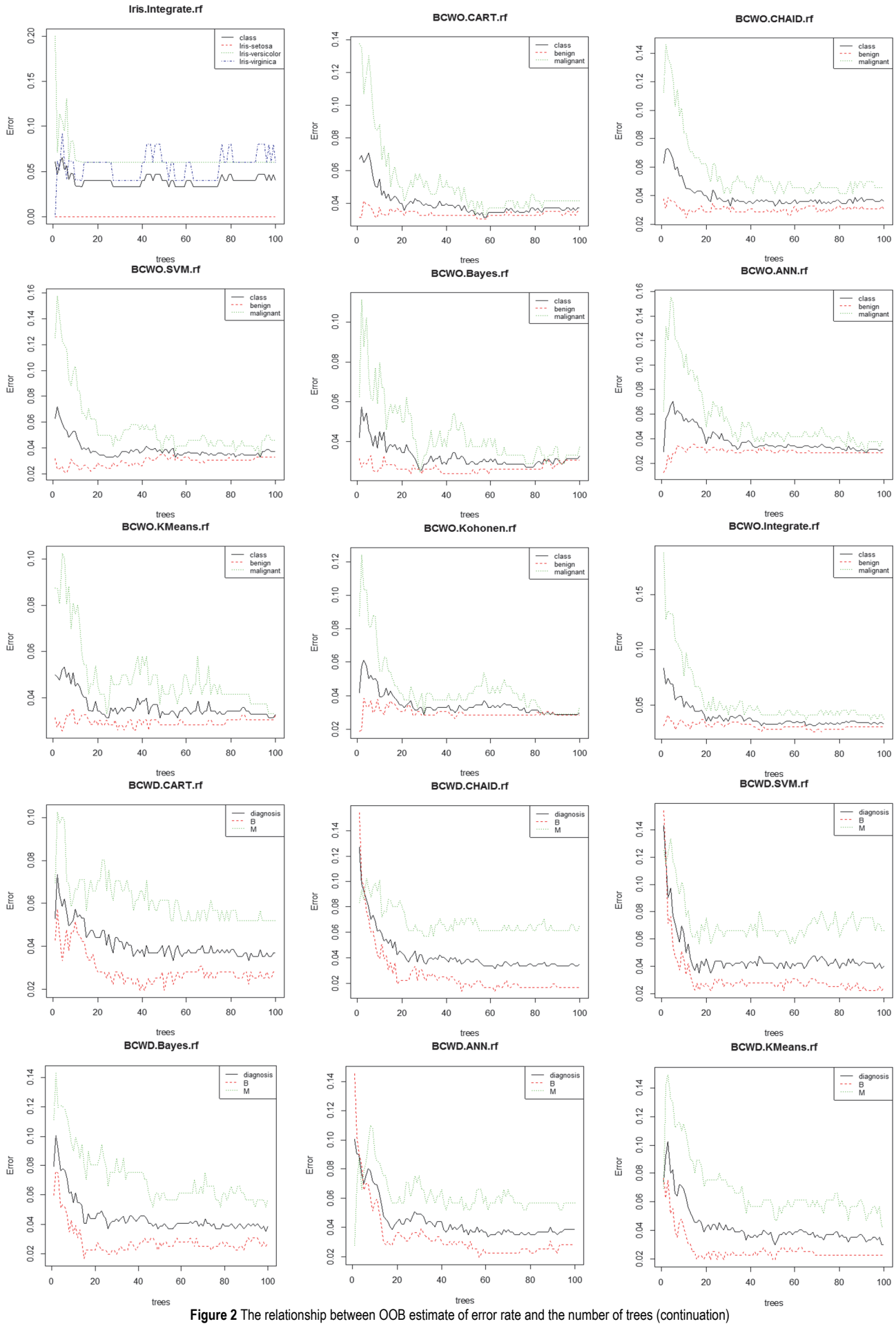


Figure 2 The relationship between OOB estimate of error rate and the number of trees (continuation)

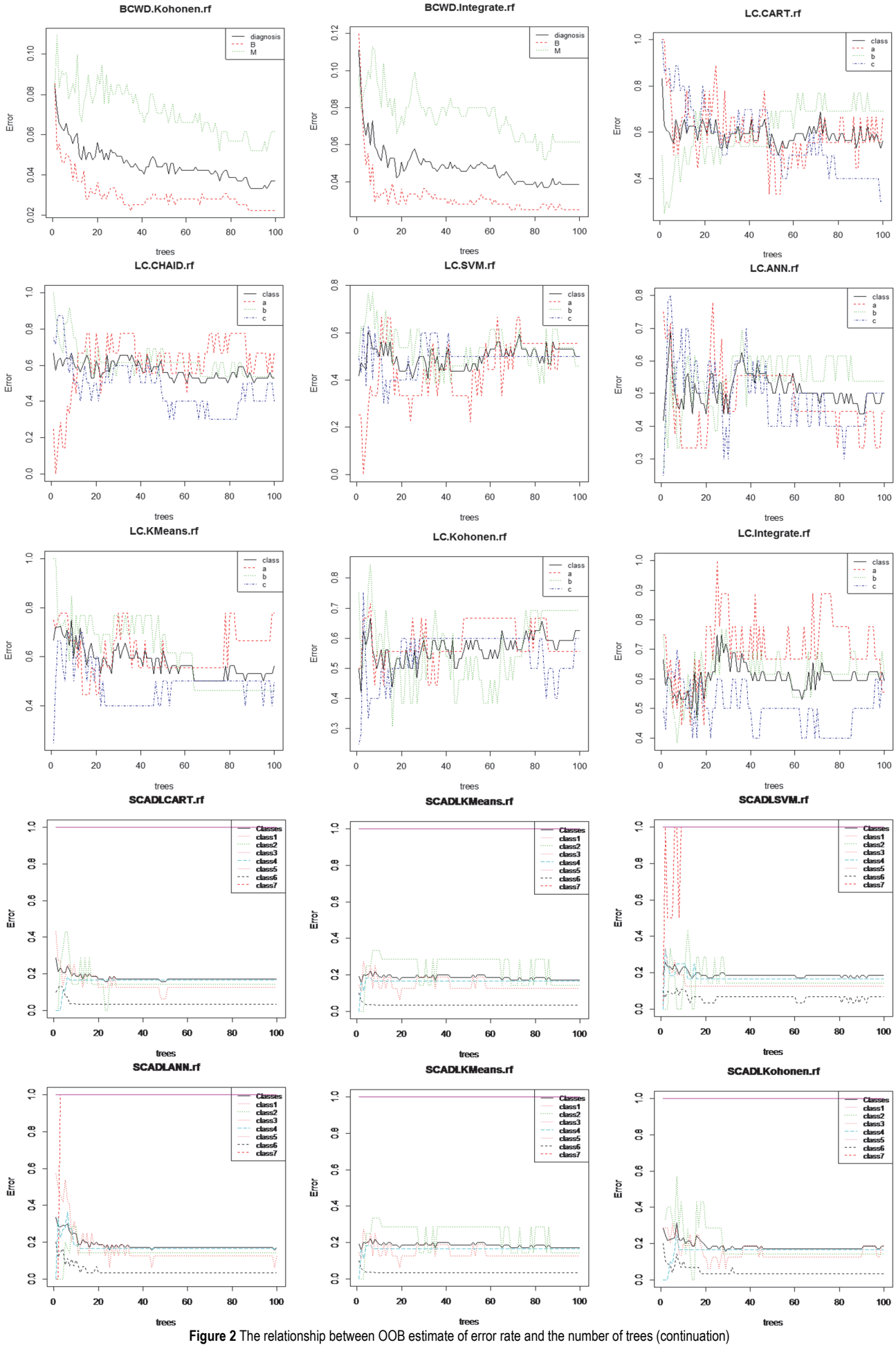


Figure 2 The relationship between OOB estimate of error rate and the number of trees (continuation)

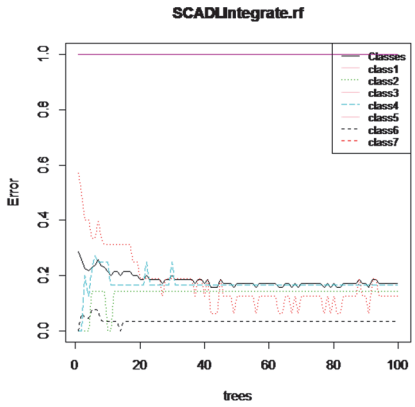


Figure 2 The relationship between OOB estimate of error rate and the number of trees (continuation)

In Fig. 2, there are a total of 46 sub-graphs. The first 4 sets of data sets each has 8 sub-graphs, and the last 2 sets of data sets each has 7 sub-graphs. The sub-graphs (aa) have the title of Lenses. CART.rf means that CART method obtains the feature importance ranking of the Lenses data and reorders the features in the Lenses data table, and then uses random forests to classify the new data table; the abscissa is the number of decision trees, and the ordinate is the OOB error estimation rate; The dotted line in the sub-picture represents the change trend of OOB error estimation rate for each category over the number of decision trees; the black solid line represents the change trend of overall OOB error estimates over the number of decision trees. From Fig. 2, it can be concluded that the OOB error estimate of the random forests decreases and then gradually balances over the number of decision trees, and we can get the number of decision trees for the best model of each data set as follows: The Lens dataset is 25, the Iris dataset is 25, the Breast Cancer Wisconsin (Original) dataset is 50, the Breast Cancer Wisconsin (Diagnostic) dataset is 50, the lung-cancer dataset is 50, and the SCADI dataset is 100.

4.4 OOB Estimate of Error Rate Analysis

Analysing each sub-graph in Fig. 3, we can know, as features were introduced sequentially, the OOB error estimation rate of the random forests decreases and then balances, indicating that the above eight feature importance evaluation methods are all effective. Comparing all methods, it is found that Integrate-RF method is more stable.

In order to further compare the above eight methods, it calculates the minimum value, average value and variance of the OOB error estimation rate of each data set in the above experiments. The results are shown in Fig. 3. Each data set uses different methods to obtain the number of features at the minimum OOB error, the results are shown in Tab. 3.

Combining Fig. 3, Tab. 3 and Tab. 4, we can know: In Lenses, there is a total of 4 features, CART-RF, SVM-RF, and BN-RF introduce the first three features to get the minimum OOB error rate 16.67%; K-Means-RF introduces the first four features to get the minimum OOB error rate of 25%; Integrate-RF introduces the first two features to get the minimum OOB error rate of 16.67%.

In Iris, there is a total of four features, CART-RF, BN-RF, K-Means-RF, Kohonen-RF. Integrate-RF introduces the first three features to get the minimum OOB error rate 4.00%; CHIAD-RF and SVM-RF introduce the first four features to get the minimum OOB error rate 3.33%; NN-RF introduce the first four features to get the minimum OOB error rate 5.33%.

In BCWO, there is a total of nine features, CART-RF introduces the first seven features to get the smallest OOB error rate being 3.43%; CHIAD-RF introduces the first nine features to get the smallest OOB error rate to be 3.58%; SVM-RF introduces the first eight features to get the smallest OOB error rate of 3.00%; BN-RF introduces the first nine features to get the smallest OOB error rate of 3.43%; NN-RF introduces the first nine features to get the smallest OOB error rate being 2.86%; K-Means-RF introduces the first nine features to get the smallest OOB error rate of 3.29%; Kohonen-RF introduces the first seven features to get the smallest OOB error rate of 3.29%; Integrate-RF introduces the first seven features to get the smallest OOB error rate to be 3.00%.

In BCWD, there is a total of thirty features, CART-RF introduces the first thirty features to get the smallest OOB error rate is 3.69%; CHIAD-RF introduces the first eight features to get the smallest OOB error rate is 3.16%; SVM-RF introduces the first twenty-five features to get the smallest OOB error rate is 2.81%; BN-RF introduced the first fourteen features to get the smallest OOB error rate is 2.99%; NN-RF introduced the first twenty-four features to get the smallest OOB error rate is 3.69%; K-Means-RF introduced the first thirty features to get the smallest OOB error rate is 2.99%; Kohonen-RF introduces the first twenty-nine features to get the minimum OOB error rate is 3.34%; Integrate-RF introduce the first twenty features to get the minimum OOB error rate is 3.69%.

In lung-cancer, there is a total of fifty-six features, CART-RF introduces the first three features to get the minimum OOB error rate of 31.25%; CHIAD-RF introduce the first two features to get the minimum OOB error rate being 28.12%; SVM-RF introduce the first twenty features to get the minimum OOB error rate of 31.25%; NN-RF introduce the first six features to get the minimum OOB error rate is 34.38%; K-Means-RF introduce the first thirty-three features to get the minimum OOB error rate is 40.62%; Kohonen-RF introduce the first thirty features to get the minimum OOB error rate is 37.50%; Integrate-RF introduce the first nine features to get the minimum OOB error rate is 28.12%.

In SCADI, there is a total of 205 features, CART-RF introduces the first four features to get the smallest OOB

error rate is 12.86%; CHIAD-RF introduce the first four features to get the smallest OOB error rate is 14.29%; SVM-RF introduce the first a hundred and thirteen features to get the smallest OOB error rate is 14.29%; NN-RF introduce the first one hundred and twenty six features to get the smallest OOB error rate of 14.29%; K-Means-RF introduce the first forty-one features to get the smallest OOB error rate of 12.86%; Kohonen-RF introduce the first fifteen features to get the smallest OOB error rate of 14.29%; Integrate-RF introduced the first 27 features to get the minimum OOB error rate is 14.29%.

From the analysis of the above results, three points can be obtained. First, eight feature importance evaluation methods are feasible and effective, especially when there are more feature amounts in the data set, the effect is more obvious. Second, Integrate-RF performs better on all data sets, indicating that Integrate-RF is more adaptable. Third, the same features are ordered differently in the data table, and the results produced are also different, indicating that the random forests extraction features in the R software package are related to the ordering of features in the data table.

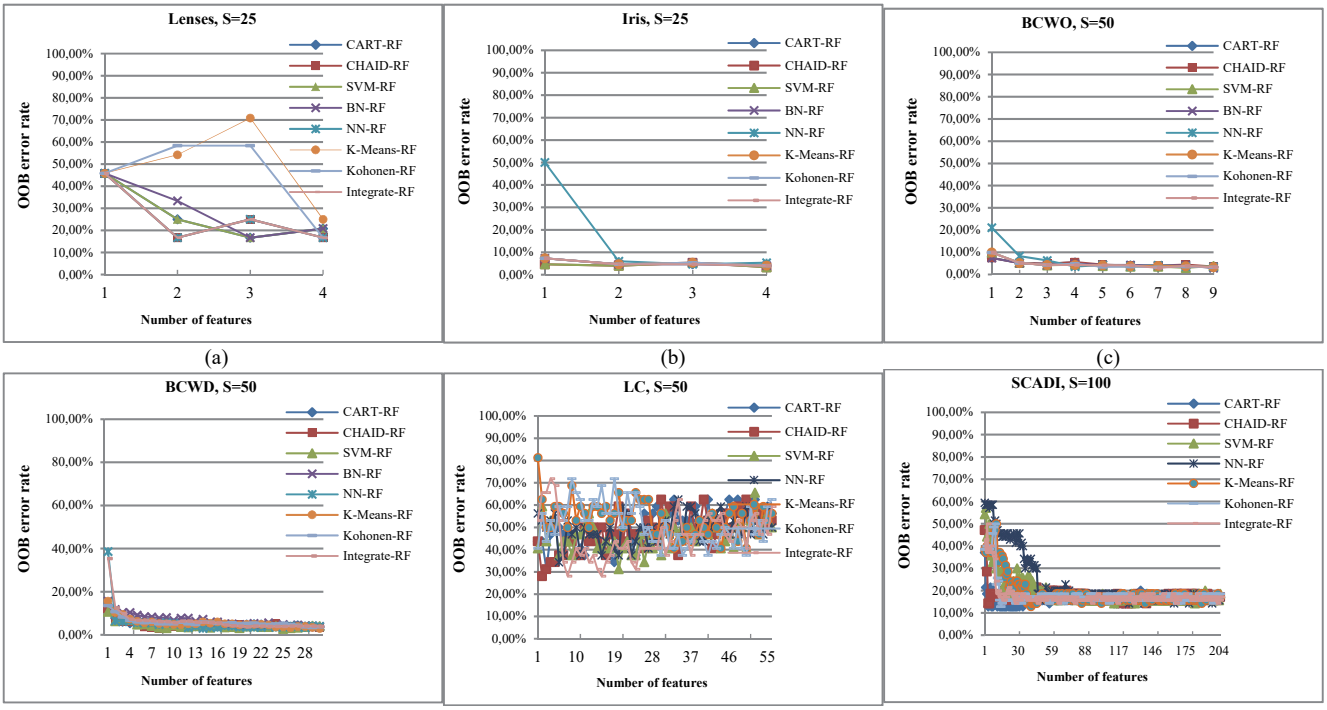


Figure 3 The relationship between OOB estimate of error rate and the number of selected features using eight feature-importance-evaluation and random forests classification methods (Among them, BCWO represents the data set Breast Cancer Wisconsin (Original), BCWD represents the data set Breast Cancer Wisconsin (Diagnostic), and LC represents the data set lung-cancer.)

Table 3 The value of OOB estimate of error rate of different features set sizes using eight feature-importance-evaluation and random forests (Min / %, AV + V / %)

	Lenses		Iris		BCWO		BCWD		lung-cancer		SCADI	
	Min	AV ± V	Min	AV ± V	Min	AV ± V	Min	AV ± V	Min	AV ± V	Min	AV ± V
CART-RF	**16.67**	27.08 ± 1.68	4.00	5.17 ± 0.02	3.43	**4.59** ± 0.02	3.69	5.12 ± 0.03	31.25	49.95 ± 0.64	**12.86**	**16.09** ± 0.05
CHIAD-RF	**16.67**	**26.04** ± 1.89	**3.33**	**4.33** ± 0.01	3.58	4.83 ± 0.01	3.16	4.52 ± 0.03	**28.12**	47.99 ± 0.68	14.29	17.75 ± 0.12
SVM-RF	**16.67**	27.08 ± 1.68	**3.33**	**4.33** ± 0.01	3.00	4.62 ± 0.04	2.81	**4.33** ± 0.02	31.25	**45.98** ± 0.43	14.29	19.35 ± 0.41
BN-RF	**16.67**	29.17 ± 1.74	4.00	5.33 ± 0.02	3.43	6.49 ± 0.32	**2.99**	5.52 ± 0.40				
NN-RF	**16.67**	26.04 ± 1.89	5.33	16.50 ± 4.99	**2.86**	**4.59** ± 0.02	3.69	6.71 ± 0.08	34.38	49.44 ± 0.47	14.29	23.17 ± 1.44
K-Means-RF	25.00	48.96 ± 3.63	4.00	5.33 ± 0.02	3.29	4.70 ± 0.04	**2.99**	5.58 ± 0.06	40.62	53.96 ± 0.59	**12.86**	19.15 ± 0.43
Kohonen-RF	**16.67**	44.79 ± 3.86	4.00	5.33 ± 0.02	3.29	4.59 ± 0.04	3.34	5.71 ± 0.04	37.50	51.23 ± 0.87	14.29	18.02 ± 0.32
Integrate-RF	**16.67**	**26.04** ± 1.89	4.00	5.17 ± 0.02	3.00	4.67 ± 0.04	3.69	6.47 ± 0.33	**28.12**	46.26 ± 0.99	14.29	17.66 ± 0.24

Note: lung-cancer and SCADI have many features, so BN-RF cannot get results.

Table 4 Number of convergent features using eight feature-importance-evaluation and random forests classification methods

	Lenses	Iris	BCWO	BCWD	lung-cancer	SCADI
CART-RF	3	4	7	30	3	4
CHIAD-RF	4	4	9	8	2	4
SVM-RF	3	4	8	25	20	113
BN-RF	3	4	9	14		
NN-RF	4	4	9	24	6	126
K-Means-RF	4	4	9	30	33	41
Kohonen-RF	4	4	7	29	30	15
Integrate-RF	2	4	7	20	9	27

Note: lung-cancer and SCADI have many features, so BN-RF cannot get results.

5 CONCLUSION

In order to overcome the dimensional disaster and over-fitting problems, and improve the efficiency of data analysis the paper proposes eight feature selection methods, they are CART-RF, CHAID-RF, SVM-RF, BN-RF, NN-RF, K-Means-RF, Kohonen-RF and Integrate-RF. The first seven methods use CART, CHAID, SVM, BN, NN, K-Means or Kohonen to evaluate feature importance, and then use forward search strategy to generate and update feature subsets; finally, use random forests to evaluate feature subsets. The eighth feature selection method uses hybrid integration of difference models and random forests. Through 6 sets of UCI data experiments, the results show that the eight methods can effectively select features, reduce the data dimension.

6 DECLARATION

(1) We note that a shorter conference version of this paper appeared in 2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI) 27. - 29. August 2021, Shanghai, China. This manuscript supplements and expands the following contents in more detail: seven basic feature selection methods structure, feature importance ranking, the relationship between OOBEER estimate of error rate and the number of trees, the value of OOB estimate of error rate of different features set sizes using eight feature-importance-evaluation and random forests classification methods, and etc.

(2) The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

(3) Data are from UCI database (http://archive.ics.uci.edu/ml/datasets.html).

(4) This work uses the software of IBM SPSS modeler 14.1 to get the feature importance ranking, and then uses the R programming to perform feature subset generation, evaluation, and selection.

7 ACKNOWLEDGMENT

This work thanks Bin Nie, Yuwen Du, Huan Li, Jianqiang Du, Yufeng Chen.

8 REFERENCES

[1] Wang, C. Z., Hu, Q. H., Wang, X. Z., Chen, D. G., Qian, Y. H., & Dong, Z. (2018). Feature selection based on neighbour hood discrimination Index. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(7), 2986-2999. https://doi.org/10.1109/TNNLS.2017.2710422

[2] Yang, Y. Y., Chen, D. G., & Wang, H. (2017). Active sample selection based Incremental Algorithm for Attribute Reduction with Rough Sets. *IEEE Transactions on Fuzzy Systems*, *25*(4), 825-838. https://doi.org/10.1109/TFUZZ.2016.2581186

[3] Ma, W. P., Zhou, X. B., Zhu, H., Li, L. W., & Jiao, L. C. (2021). A two-stage hybrid ant colony optimization for high-dimensional feature selection. *Pattern Recognition*, *116*(1). https://doi.org/10.1016/j.patcog.2021.107933

[4] Jain, D. & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, *19*(3), 179-189. https://doi.org/10.1016/j.eij.2018.03.002

[5] Wang, X. H., He, Y. D., & Wang, L. Z. (2018). Study on mutual information and fractal Dimension-Based unsupervised feature parameters selection: application in UAVs. *Entropy*, *20*(9), 674. https://doi.org/10.3390/e20090674

[6] Bhadra, T. & Bandyopadhyay, S. (2021). Supervised feature selection using integration of densest subgraph finding with floating forward-backward search. *Information Sciences*, *566*(1-18). https://doi.org/10.1016/j.ins.2021.02.034

[7] Gan, J. Z., Wen, G. Q., Yu, H., Zheng, W., & Lei, C. (2020). Supervised feature selection by self-paced learning regression. *Pattern Recognition Letters*, *132*(30-37). https://doi.org/10.1016/j.patrec.2018.08.029

[8] Kundu, P. P. & Mitra, S. (2017). Feature selection through message passing. *IEEE Transactions on Cybernetics*, *47*(12), 4356-4366. https://doi.org/10.1109/TCYB.2016.2609408

[9] Pino Angulo, A. & Shin, K. (2017). Improving Classification Accuracy by Means of the Sliding Window Method in Consistency-Based Feature Selection. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-319-67786-6_12

[10] Yang, R. T., Zhang, C. J., Gao, R., & Zhang, L. N. (2016). A novel feature extraction method with Feature Selection to identify Golgi-Resident Protein Types from Imbalanced Data. *International Journal of Molecular Sciences*, *17* (2). https://doi.org/10.3390/ijms17020218

[11] Sluga, D. & Lotric, U. (2017). Quadratic Mutual information feature selection. *Entropy*, *19*(4), 157-173. https://doi.org/10.3390/e19040157

[12] Abad, J. M. N. & Soleimani, A. (2018). Novel feature selection algorithm for thermal prediction model. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *26*(10), 1831-1844. https://doi.org/10.1109/TVLSI.2018.2841318

[13] Bradley, P. E., Keller, S., & Weinmann, M. (2018). Unsupervised feature selection based on ultrametricity and sparse training data: A Case Study for the Classification of High-Dimensional Hyperspectral Data. *Remote Sensing*, *10*(10), 1564. https://doi.org/10.3390/rs10101564

[14] Lu, M. (2019). Embedded feature selection accounting for unknown data heterogeneity. *Expert Systems with Applications*, 119, 350-361. https://doi.org/10.1016/j.eswa.2018.11.006

[15] Zhao, J., Chen, L., Pedrycz, W., & Wang, W. (2019). Variational Inference-Based Automatic Relevance Determination Kernel for Embedded Feature Selection of Noisy Industrial Data. *IEEE Transactions on Industrial Electronics*, *66*(1), 416-428. https://doi.org/10.1109/TIE.2018.2815997

[16] Zheng, Y. F., Li, Y., Wang, G., Chen, Y. P., Xu, Q., Fan, J. H., & Cui, X. T. (2018). A novel hybrid algorithm for feature selection. *Personal and Ubiquitous Computing*, *22*(5-6), 971-985. https://doi.org/10.1007/s00779-018-1156-z

[17] Grajski, K. A., Breiman, L., Di Prisco, G. V., & Freeman, W. J. (1987). Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART. *IEEE transactions on bio-medical engineering*, *33*(12), 1076-1086. https://doi.org/10.1109/TBME.1986.325684

[18] van Diepen, M. & Franses, P. H. (2006). Evaluating chi-squared automatic interaction detection. *Information Systems*, *31*(8), 814-831. https://doi.org/10.1016/j.is.2005.03.002

[19] Guenther, N. & Schonlau, M. (2016). Support vector machines. *Stata Journal*, *16*(4), 917-937. https://doi.org/10.1177/1536867X1601600407

[20] Darwiche, A. (2010). Bayesian Networks. *Communications of the ACM*, *53*(12), 80-90.

https://doi.org/10.1145/1859204.1859227
[21] Presnell, S. R. & Cohen, F. E. (1993). Artificial neural networks for pattern recognition in biochemical sequences. *Annual review of biophysics and biomolecular structure*, *22*(283-298). https://doi.org/10.1146/annurev.bb.22.060193.001435
[22] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451-461. https://doi.org/10.1016/S0031-3203(02)00060-2
[23] Pan, Y., Zhang, L. M., Li, Z. W. (2020). Mining event logs for knowledge discovery based on adaptive efficient fuzzy Kohonen clustering network. *Knowledge-Based Systems, 209*. https://doi.org/10.1016/j.knosys.2020.106482
[24] Wang, Z., Nie, B., Du, Y. W., Li, H., Du, J. Q., & Chen, Y. F. (2021). Feature selection using different evaluate strategy and random forests. *2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI).* 27-29 Shanghai, China. https://doi.org/10.1109/ICCEAI52939.2021.00062

Contact information:

Zhuo WANG, Master, Lecture
(Corresponding author)
School of Software,
Nanchang University,
Nanchang, Jiangxi, China
E-mail: 52373793@qq.com