2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)

Margin Based Permutation Variable Importance: a Stable Importance Measure for Random Forest

Liu Pei
Department of Automation
Xiamen University
Xiamen,Fujian Province
Email: peiliu@stu.xmu.edu.cn

Yongxuan Lai
School of software
Xiamen University
Xiamen,Fujian Province
Email: laiyx@xmu.edu.cn

Peng Piao
Department of Automation
Xiamen University
Xiamen,Fujian Province
Email: piaopeng@stu.xmu.edu.cn

Fan Yang*
Department of Automation
Xiamen University
Xiamen,Fujian Province
Email: yang@xmu.edu.cn

***Abstract*—Permutation based variable importance measure (VIM) has been widely used in various research fields. For example, in gene expression studies, it has been regarded as a screening tool to select a subset of relevant genes for subsequent analysis or better predictive performance. However, little effort has been devoted to the stability of variable importance measures. In this paper, margin based permutation variable importance measures (VIM-MDs) are proposed, which utilize the similarity between margin distribution before and after random permutation to evaluate the importance of variables. Experiments on six benchmark datasets show that the VIM-MDs outperform permutation based variable importance measure in terms of both global stability and predictive accuracy, which indicates that the proposed method could be used as an effective and stable variable importance measure for random forest.**

Keywords-random forest; variable importance; stable feature selection;margin distribution

I. INTRODUCTION

In bioinformatics, genomics, and related fields, thousands of genes are investigated. In genetic data, the number of variables tremendously surpasses the number of observations, while not all of the variables are relevant for prediction. Further, some irrelevant variables have a negative impact on the model accuracy, so identifying the relevant variables is a critical task for prediction. Feature selection, also known as variable selection, is an effective way to rank all features and identify a subset of relevant genes, which removes the redundant features and provides a valid feature subset for prediction. Random forest, owing to its ability to tackle high-dimensional data and select relevant features, has been regarded as a screening tool for gene expression studies. Lunetta et al. [1] found that genetic-relevant variables can be identified more efficiently by means of random forests than other screening methods.

Random forest is an ensemble of classification or regression trees. The data used to build each tree is sampled without replacement from the original training data, and at each split the candidate variables are a random subset of all the variables. In order to get low biased trees, each tree grows fully. At the same time, bagging and randomly splitting gives rise to low relationship of the individual trees. Random forest achieves high prediction accuracy even for high-dimensional data with correlated and redundant features, which is common in many fields, e.g. gene expression data. Its inbuilt variable importance measures has been becoming one of the most frequently-used feature selection methods in these fields. For example, in genome-wide association studies (GWAS) random forests are used to detect gene-gene interactions [1].

Two variable importance measures are widely used, i.e. permutation importance and Gini importance. Unfortunately, some studies showed that both importance measures were biased. Strobl et al. [2] found that Gini importance was biased towards the categorical variables which had many distinct values while permutation importance showed a bias towards correlated variables. To solve these problems, several unbiased algorithms have been proposed. Strobl et al. [2] introduced an alternative implementation of random forest which provides unbiased selection in the individual classification trees. Strobl et al. [3] developed a new conditional permutation scheme for the computation of the variable importance measure. Sophia S.F. Lee et al. [4] proposed an EM-random forest and new variable importance measures based on Haseman-Elston quantitative trait linkage analysis. Andr Altmann et al. [5] introduced a heuristic to normalize feature importance measures to correct the variable importance bias in favor of categorical variables with a large number of categories. Baptiste Gregorutti et

978-1-5386-1829-5/17/$31.00

al. [6] provided some theoretical insights on the effect of the corrections on the permutation importance measure and compare recursive and non-recursive approaches through an extensive simulation study on several classification and regression tasks to illustrate the efficiency of the recursive feature elimination (RFE) algorithm for selecting a small number of variables together with a good prediction error.

However, little effort has been devoted to the stability of variable importance measures. The stability of a feature selection algorithm is the robustness of the feature preferences to changes in training sets, and stability quantifies how different training sets affect the feature preferences [7]. Alexandros Kalousis et al. [7] firstly proposed a framework that measures the stability of feature selection algorithms, and examined three different stability measures and proposed a resampling technique to empirically estimate them. Recently, the stability of random forest variable importance measures has attracted attentions to the bioinformatics community. There are two VIMs from random forest[8] : Mean Decrease Gini (MDG), the average across the forest of the decrease in Gini impurity for a variable, and Mean Decrease Accuracy (MDA), the average across the forest of the accuracy for the variable minus the decrease in accuracy after random permutation of the variable. M. Luz Calle and Vłctor Urrea [9] emphasized the value of exploring ranking stability of two VIMs and showed that MDA were unstable to small perturbations of the dataset while MDG provide more robust results. However, Kristin [10] examined data-specific characteristics on ranking stability of VIMs and showed that rankings based on the MDG measure is sensitive to within-predictor correlation and differences in category frequencies while MDA is more robust. Wang et al. [11] introduced a new concept of intrinsic stability of VIMs to concern the influence of intrinsic randomness in algorithm design. The goal of this article is to design a more stable VIM for random forest, which is based on the concept of margin distribution. Margin is an important concept to understand the working mechanism of random forest and other ensemble methods. Leo Breimain [8] defined the margin of a training example as the proportion of votes for the correct class minus maximum proportion of votes for other classes. In this work, we utilize the similarity between two margin distributions before and after randomly permuting each variable to measure the variable importance for each variable. We examine the stability of the proposed methods (VIM-MDs) in comparison with VIM as well as the predictive accuracy. We evaluate the stability of both VIM and VIM-MDs based on the similarity of the rankings and the selected feature subsets with respect to disturbance in training set.

This rest of the paper is organized as follows: in Section II we introduce the variable importance measures and the concept of stable feature selection. In Section III, the proposed margin based permutation variable importance is presented. Experiments and discussion are given in section IV. Section V concludes the paper.

II. Related work

A. Variable importance measures

1) Gini importance: The average decreases on the Gini impurity over all splits yields the Gini variable importance measure (MDG). However, MDG is prone to bias when categorical predictors have too many distinct values [2] and it also leads to undesirable feature ranking when handling imbalanced data [10]. In this paper we mainly focus on the following permutation based variable importance.

2) Permutation importance: In this paper, the abbreviation 'VIM' denotes the permutation based VIM (MDA) which evaluates the variable importance with mean decrease in the OOB (out-of-bag) accuracy of decision trees caused by random permutation of the variables. Let $\{h_1, h_2, \cdots, h_{tree}\}$ be an ensemble of decision trees and OOB_i be the out-of-bag examples of $i-th$ tree, the importance score of the variable m is then defined as:

$$VIM(m) = \frac{1}{ntree} \sum_{i=1}^{ntree} \frac{\sum_{j \in OOB_i} I(y_i = h_i(x_j)) - \sum_{j \in OOB_i} I(y_i = h_i(x_j^m))}{|OOB_i|} \tag{1}$$

where y_j is the true label of the example j, $h_i(x_j)$ and $h_i\left(x_j^m\right)$ are the predicted label before and after random permutation of variable m respectively, and I is the indicator function.

The reason that Breiman used the mean decrease in OOB accuracy of decision trees rather than the decrease in OOB accuracy of the forest may lie in the fact that the former is more sensitive than the latter when randomly permuting a variable. In more details, the OOB estimates of random forest may not change at all when randomly permuting a variable, especially on the high-dimensional data since there are lots of correlated variables. However, even MDA may be unreliable and unstable as a feature selection method as Strobl et al. [2] pointed out. Using permutation test, Andr Altmann et al. [5] introduced a heuristic to normalize feature importance measures to correct the bias in favor of categorical variables with many categories. Further, M. Luz Calle and Vłctor Urrea [9] and Kristin [10] also emphasized the value of exploring ranking stability of two VIMs and examined the stability of MDA in different situations.

B. Definition of margin

In our previous work [12], we found that the change in the mean sample margin in random forest when permuting a variable is more sensitive than the OOB accuracy of the decision trees, and proposed margin based variable importance. In this article, we further use the margin distribution to design new permutation importance measure. Let $h_1(X), h_2(X), \cdots, h_k(X)$ be an ensemble of classifiers,

and with the training set drawn randomly from the distribution of the random vector (X, Y), the margin function is defined as:

$$mg(X,Y) = av_k I(h_k(x) = Y) - \max_{j \neq Y} av_k I(h_k(x) = j) \tag{2}$$

where $I(\bullet)$ denotes the indicator function, and $av_k(\bullet)$ denotes the mean operation [8]. The margin of an instance is defined as the proportion of votes for the correct class minus maximum proportion of votes for the other classes. Thus, positive margin for an instance indicates correct classification, and vice versa. The larger the margin is, the more confidence in the classification. The generalization error is given by

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{3}$$

C. Stable feature selection

1) The stability issue of feature selection: The stability of feature selection algorithm is defined as the sensitivity of feature ranking to changes in the training set [7]. Exploring the stability aims at demonstrating that the feature ranking is robust to random disturbance in training sample, and hence reliable for some feature selection tasks, e.g. biomarker discovery.

2) The stability measures: Measuring stability needs similarity measure for feature preferences. There are three ways to measure the similarity. The first one is based on the weight or score which indicates the importance of each feature, and the second one is based on the ranking which is the simplification of the first one. The last one is based on the selected feature subsets where no score or ranking is considered. We use the latter two methods in this paper.

Let $(f_1, f_2, \cdots, f_m)$be the feature set, a corresponding ranking vector can be represented as,

$$R = (r_1, r_2, ..., r_m), 1 \leq r_i \leq m, r_i \in N^+ \tag{4}$$

to measure similarity between two ranking vectors r_i and r_i^*, we use *Spearman's Rank Correlation Coefficient(SRCC)* assuming that all the ranks are distinct integers [9],

$$SRCC(R, R^*) = 1 - 6\sum_i^m \frac{(r_i - r_i^*)}{m(m^2-1)} \tag{5}$$

The range of $SRCC$ is [-1,1], where 1 means two ranking vectors are identical while -1 means they are fully opposed.

Given $(f_1, f_2, \cdots, f_m)$, a feature subset vector can be defined as,

$$S = (s_1, s_2, ..., s_i), 0 < i \leq m \tag{6}$$

and we use the Jaccard index [9] and Kuncheva index [13] to measure similarity between two feature subset vectors with the same length,

$$JI(S, S^*) = \frac{|S \cap S^*|}{|S \cup S^*|} \tag{7}$$

The range of Jaccard index is [0,1], where 0 means two feature subset vectors are completely different while 1 means they are identical. And Kuncheva's index is defined as,

$$KI(S, S^*) = \frac{|S \cap S^*|m - k^2}{k(m-k)} \tag{8}$$

where k is the size of feature subset.

III. Margin based permutation variable importance measure

As shown in above section, theoretically, compared with OOB accuracy, margin plays a more essential role in the generalization ability of random forest. In our previous studies[12] the performance of using average margin of training examples for feature ranking has been investigated. However, average margin only gives partial statistical information. In this section we propose to utilize margin distribution on the training examples to evaluate the feature importance (VIM-MDs). Specifically, in order to compare to distributions, we use the widely used *Cosine similarity*, *Pearson Correlation Coefficient (PCC)* and *Spearman's Rank Correlation Coefficient (SRCC)* respectively to calculate the similarity between two margin distribution before and after random permutation of each variable. Obviously, the more similar the two distribution is, the less important the variable is.

Let $Z = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ be the training set and $f = \{h_1, h_2, \cdots, h_m\}$ be an ensemble of decision trees, and $M = \{m_1, m_2, \cdots, m_n\}$ be the margin vector for instances in training set. Randomly permute a variable *i*, and the corresponding margin vector can be represented as,

$$M^i = \{m_1^i, m_2^i, ..., m_n^i\} \tag{9}$$

Cosine similarity. VIM-MDI measures similarity between two margin vectors M and M^i with Cosine similarity [14]:

$$Cosine(M, M^i) = \frac{M \cdot M^i}{||M||_2||M^i||_2} = \frac{\sum_{j=1}^n M_j M_j^i}{\sqrt{\sum_{j=1}^n M_j^2}\sqrt{\sum_{j=1}^n M_j^i}} \tag{10}$$

where M_j and M_j^i are components of vector M and M^i respectively.

Pearson Correlation Coefficient. VIM-MDII measures similarity between two margin vectors M and M^i with Pearson Correlation Coefficient [9]:

$$PCC(M, M^i) = \frac{\sum_j (m_j - \mu_M)(m_j^i - \mu_M^i)}{\sqrt{\sum_j (m_j - \mu_M)^2 \sum_j (m_j^i - \mu_M^i)^2}} \tag{11}$$

where the μ_M and μ_M^i are the mean margins respectively. It has a value between -1 and 1, where -1 is total negative linear correlation, 0 is no linear correlation, and 1 is total positive linear correlation.

Similarly, VIM-MDIII is implemented using Spearman's Rank Correlation Coefficient(SRCC).

The algorithm is described as follows. Given the original margin vector $M = \{m_1, m_2, \cdots, m_n\}$, for a variable *i*,

1. Randomly permute variable *i* and get a new training set, then put the new data down the trees and get a new margin vector,

$$M^i = \{m_1^i, m_2^i, ..., m_n^i\} \tag{12}$$

2. Compute the similarity between two margin vectors before and after permutation, for VIM-MDI,

$$S_1^i = Cosine(M, M^i) \tag{13}$$

for VIM-MDII,

$$S_2^i = PCC(M, M^i) \tag{14}$$

for VIM-MDIII,

$$S_3^i = SRCC(M, M^i) \tag{15}$$

3. Get the importance scores for each variable using VIM-MDs.

IV. Experiments and Results

A. Datasets

Experiments were conducted on three high dimensional datasets and three low dimensional datasets. The datasets used is shown in TABLE I, where n stands for the size of the datasets, p denotes the number of features, and m represents the number of classes.

Table I
DATASETS USED IN EXPERIMENTS.

Date sets	n	p	m
Colon	62	2000	2
Leukemia	72	5147	2
Central Nervous System outcome	60	7129	2
p-gene	106	57	2
SPECT-HEART	80	44	2
hearts	267	44	2

B. Experiment design

In order to compare the proposed method with permutation importance (VIM) in terms of stability and predictive accuracy, four experiments were conducted. For computing the variable importance, we use the RF with 500,1000,and 2000 trees respectively in a 10-fold cross-validation setting.

1) Experiment I: In this experiment, our main purpose is to observe the margin distribution before and after random permutation of a variable:

1.Get the margin distribution before permutation using the original data.

2.Permute a variable two times and in each permutation get a margin distribution.

3.Enumerate all the variables, and repeat the above steps.

2) Experiment II: In this experiment, we compare the sensitivity of OOB error of random forest, mean OOB error of decision trees in random forest, and the margin distribution to random permutation:

1.Get the OOB error of random forest, mean OOB error of decision trees and margin distribution with the original data.

2.Permute a variable ten times, and in each permutation get a new OOB error of random forest, a new mean OOB error of decision trees, and new margin vectors for all instances. Then calculate the rates of change in OOB error, in mean OOB error of decision trees(RCOEDT), and in margin(RCM) respectively before and after each permutation.

3.Calculate the mean and standard deviation of the rates.

4.Enumerate all the variables, and repeat the above steps.

3) Experiment III: In this experiment, we calculate the importance scores to examine the stability of VIM-MD and VIM with respect to random permutation.

1.Get the OOB error of decision trees and margin distribution before permutation using the original data.

2.Permute a variable ten times, in each permutation get a new OOB error of decision trees and a new margin distribution, then calculate the changes in OOB error of decision trees, and calculate the similarity of two margin distribution before and after random permutation with Cosine Similarity, Pearson Correlation Coefficient and Spearman's Rank Correlation Coefficient respectively, i.e. the importance scores of VIM and VIM-MDs.

3.Enumerate all the variables, and repeat the above steps.

4) Experiment IV: The goal of this experiment is to compare the performance of VIM-MDs with VIM in terms of both stability and accuracy. Three margin based permutation importance methods, VIM-MDs as well as VIM were performed and the stability of the results of feature rankings were evaluated with Spearman's Rank Correlation Coefficient, Jaccard index and Kuncheva index respectively.

1. To get different training sets, 10-fold cross-validation was executed, and results of ranked feature list of the four algorithms(i.e. VIM and VIM-MDs) were recorded for each fold.

2. The similarity of each pair of ranked feature list, a total of $10(10-1)/2$ pairs, is calculated using the three evaluation metrics. For Spearman's Rank Correlation Coefficient, it was conducted on the total rankings while for Jaccard index and Kuncheva index it was performed on different given numbers of selected features. The stability scores were then averaged over all pairs.

3. For each fold, execute a random forest model with a given size of selected feature subsets. The average predictive accuracies were then recorded for each size.

C. Results and discussion

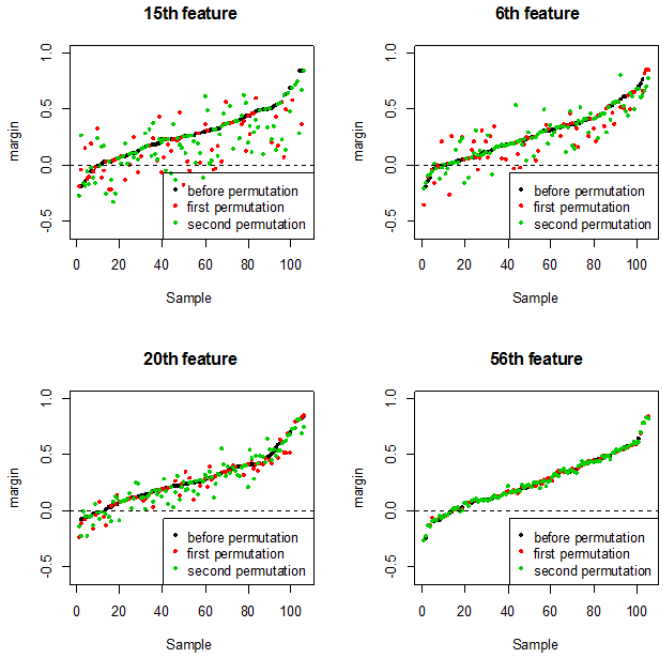

Figure 1. Margin distribution before and after permuting a variable(on p-gene data).

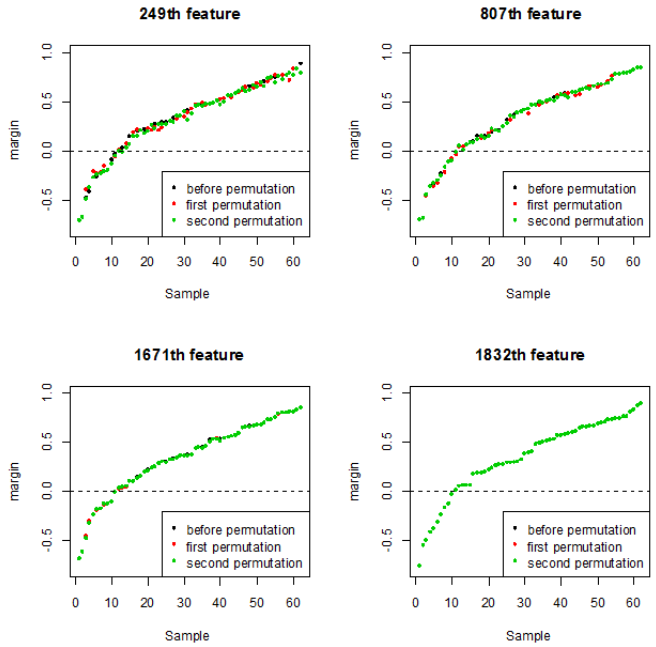

Figure 2. Margin distribution before and after permuting a variable (on Colon data).

Table II
RESULTS OF MEAN MARGIN AND VARIATION OF MEAN MARGIN.

Datasets	Features	Mean Margin			Variation of Mean Margin	
		Before Permutatio	First Permutatio	Second Permutation	BP&FP	BP&SP
p-gene	15th	0.2749472	0.1968627	0.2098289	0.07808446	0.0651183
	6th	0.2768138	0.254727	0.2586456	0.0220868	0.0181682
	20th	0.2742814	0.2614707	0.2577981	0.0128107	0.0164833
	56th	0.2599312	0.2598325	0.259658	0.0000987	0.0002732
Colon	249th	0.3374366	0.3263435	0.3186269	0.0110931	0.0188097
	1423th	0.3382442	0.3299134	0.333181	0.0083308	0.0050632
	807th	0.3446062	0.3428854	0.3431606	0.0017208	0.0014456
	1832th	0.3470133	0.3470133	0.3470133	0	0

1) Experiment I: For the sake of visualization, several examples of margin distribution before and after random permutation of a feature are displayed in Fig.1 (four features of p-gene data($p = 56$)) and Fig.2 (four features of Colon data($p = 2000$)) respectively. Fig.1 shows that the 56th feature is non-informative, because the margin distribution before and after permutation of that feature are almost the same. The most informative variable is the 15th feature, since the changes in margin distribution is the greatest, followed by the 6th feature and the 20th feature. Generally speaking, the margin distribution get worse after permutation, the margins of most instances decreased except a small number of instances.

Compared to Fig.1, the change in margin distribution before and after feature permutation is much smaller in Fig.2. The reason may lies in the fact that Colon is a high-dimensional dataset, and random permutation of a single variable while its correlated variables remaining the same will not has significant impact on the margin distribution.

TABLE II shows the variation of mean margins. Note that the changes in mean margin is small and may not be statistically significant. This motivates us to use statistics like Pearson Correlation Coefficient to compute the similarity of two margin distributions as the important scores, rather than change in mean margin proposed in our previous work [12].

Table III
RESULTS OF THE RATES OF CHANGE IN OOB ERROR(RCOE),OOB ERROR OF DECISION TREES(RCOEDT),MARGIN(RCM) ON P-GENE.

Feature	RCOEDT		RCM		RCOE	
	mean	std	mean	std	mean	std
15	0.0831	0.0051	0.1837	0.029	**0.2143**	2.93E-17
16	0.0736	0.0053	0.1473	0.0266	**0.2**	2.93E-17
6	0.0422	0.0043	0.1044	0.0318	**0.1538**	2.93E-17
41	0.0372	0.0069	0.0831	0.0204	**0.1176**	0
20	0.0203	0.0024	0.0528	0.0193	**0.0833**	1.46E-17
39	0.0131	0.0058	0.041	0.0262	**0.0667**	0
30	0.0029	0.0035	**0.009**	0.0177	0	0
31	0.0088	0.0061	**0.0142**	0.0136	0	0
56	0.0017	0.0013	**0.0053**	0.005	0	0
38	0.0067	0.0032	**0.0225**	0.0146	0	0
11	0.0034	0.005	**0.0064**	0.0083	0	0
32	0.0026	0.0038	**0.0097**	0.0103	0	0
45	0.0039	0.0033	**0.0106**	0.0089	0	0
33	0.0044	0.0033	**0.0062**	0.0083	0	0
56	0.0064	0.0021	**0.0124**	0.007	0	0

2) Experiment II: The rates of changes in OOB error(RCOE), OOB error of decision trees(RCOEDT), and margin(RCM) on p-gene and Colon are displayed in TABLE

Table IV
RESULTS OF THE RATES OF CHANGE IN OOB ERROR(RCOE), OOB ERROR OF DECISION TREES(RCOEDT),MARGIN(RCM) ON COLON.

Feature	RCOEDT		RCM		RCOE	
	mean	std	mean	std	mean	std
249	0.0365	0.0026	**0.0532**	0.0176	0	0
1671	0.0262	0.002	**0.0271**	0.008	0	0
1772	0.0163	0.0029	**0.0168**	0.0047	0	0
493	0.0133	0.0044	**0.0327**	0.0086	0	0
513	0.0102	0.0031	**0.0167**	0.0051	0	0
765	0.0102	0.0022	**0.0151**	0.0039	0	0
245	0.0021	0.0024	**0.0054**	0.004	0	0
1423	0.0081	0.0015	**0.0125**	0.004	0	0
267	0.0077	0.0022	**0.0122**	0.0045	0	0
1771	0.0016	6.27E-04	**0.0036**	0.0014	0	0
1411	0.0026	9.63E-04	**0.0027**	0.0021	0	0
377	0.0031	0.0011	**0.0035**	0.0022	0	0
31	0.0053	9.96E-04	**0.0076**	0.0029	0	0
138	0.0031	9.65E-04	**0.0075**	0.0038	0	0
1935	0.0024	8.78E-04	**0.0048**	0.002	0	0
780	0.0055	0.0016	**0.0141**	0.0041	0	0
1668	0.0027	7.17E-04	**0.0033**	0.0021	0	0
1058	0.004	0.0021	**0.0062**	0.0025	0	0
1825	9.83E-04	5.04E-04	**0.0018**	7.76E-04	0	0
807	6.43E-04	8.79E-04	**0.0014**	0.0011	0	0
1067	3.20E-04	8.32E-04	**8.11E-04**	0.0023	0	0
972	3.27E-04	3.30E-04	**9.95E-04**	9.95E-04	0	0
60	9.18E-04	2.24E-04	**0.0023**	0.0014	0	0
33	7.30E-04	3.96E-04	**0.0016**	9.44E-04	0	0
44	1.67E-04	4.59E-04	**5.81E-04**	8.65E-04	0	0
983	0	0	0	0	0	0
1156	0	0	0	0	0	0
952	0	0	0	0	0	0
960	0	0	0	0	0	0
1015	0	0	0	0	0	0
1038	0	0	0	0	0	0
1832	0	0	0	0	0	0

III and TABLE IV respectively. Similarly, we only show partial results for some features, i.e. 15 features of p-gene(with $p = 57$) and 32 features of Colon(with $p = 2000$).

From TABLE III, for the low-dimensional data p-gene, we can find that, when the features are relatively non-informative, such as the 30th feature, 31th feature, etc., the values of RCOE are always zero, that is to say, when permuting those features, the OOB error does not change at all. However, in that case, the OOB errors of decision trees and the margin still change, and the margin changes are greater than the OOB error of decision trees. When the features are informative, such as the 15th feature, 16th feature, etc., RCOE is the largest, followed by RCM and then RCOEDT.

For the results of high-dimensional data Colon in TABLE IV, we can see that whether the feature is informative or non-informative, RCOE is always zero. While, RCM is always larger than RCOEDT, except the most non-informative features. In that case, the values of the three measures are all zero.

From the above observations, we can draw the following conclusions. First, the OOB error of random forest is extremely insensitive for high dimensional data. Actually, in permutation based variable importance, Breiman used the average OOB error of decision trees rather than the OOB error of random forest to measure the variable importance. Second, margin is more sensitive to random permutation than the mean OOB error of decision trees, which motivates us to use margin distribution to measure the importance of variables instead.

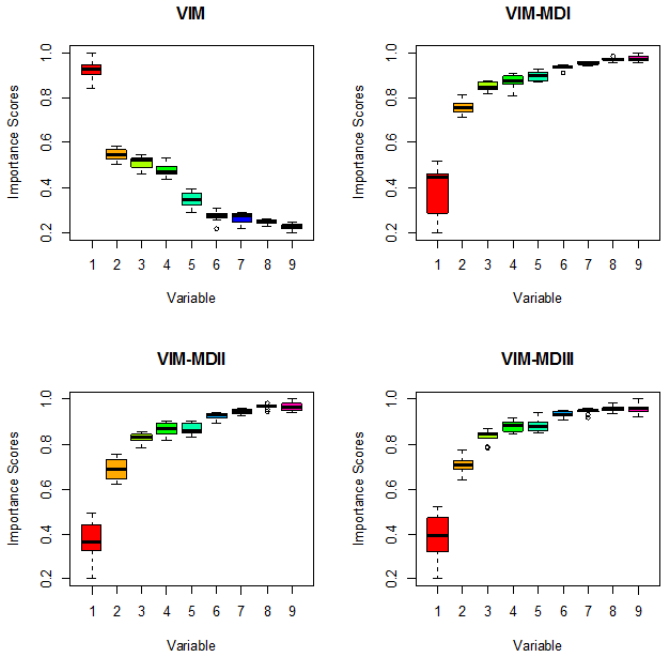

Figure 3. The importance scores of VIM and VIM-MDs on p-gene dataset.

3) Experiment III: The importance scores of top nine features selected by four VIMs are displayed in Fig.3 (on p-gene) and Fig.4 (on Colon). They are ordered according to the mean of the important scores of ten permutations, and are normalized to [0.2, 1].

Both Fig.3 and Fig.4 show that the variance of VIM scores of the top ranking feature is significantly lower than that of VIM-MDs, that is to say, the stability with respect to random permutation of the top ranking feature selected by VIM is more desirable than VIM-MDsHowever, for other features, the performances of VIM and VIM-MD are similar. It indicates that VIM-MDs can be used as an effective and robust variable importance measure for its low variance in different trials of random permutation.

4) Experiment IV: First, note that the Spearman's Coefficient provides an overall evaluation of stability by taking into account all features while the Jaccard and Kuncheva index give more details by focusing on a given size of top ranked features.

TABLE V summarizes the Spearman's Rank Correlation Coefficient of the feature rankings of VIM and VIM-MDs

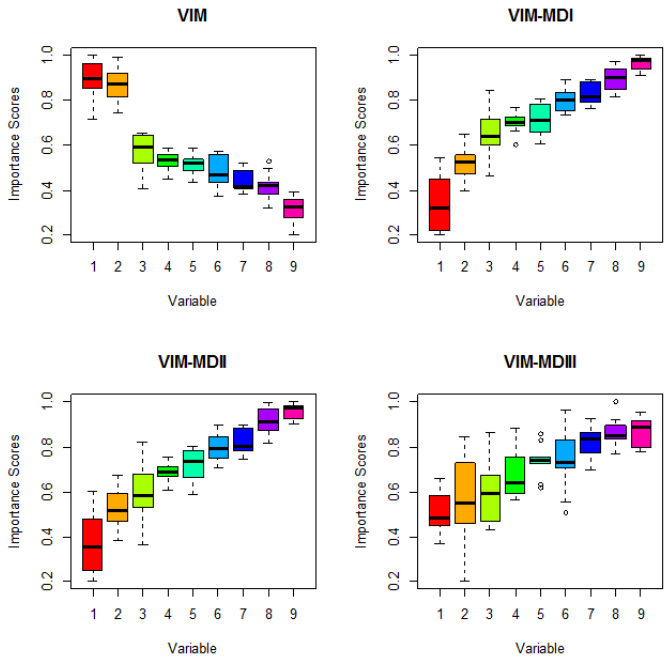

Figure 4. The importance scores of VIM and VIM-MDs on Colon dataset.

Table V
SPEARMANS RANK CORRELATION COEFFICIENT OF VIM AND VIM-MDS (NTREE=500, 1000, 2000).

Dataset	measure	Spearmans Rank Correlation Coefficient					
		ntree=500		ntree=1000		ntree=2000	
		mean	Standard deviation	mean	Standard deviation	mean	Standard deviation
colon	VIM	0.3509	0.0272	0.2934	0.0244	0.3145	0.0182
	VIM-MDI	0.5313	0.0193	0.5502	0.0214	0.6163	0.0121
	VIM-MDII	0.5302	0.0174	0.5498	0.0214	0.6159	0.0123
	VIM-MDIII	**0.5647**	0.0245	**0.5768**	0.0264	**0.5753**	0.0273
central	VIM	0.5095	0.013	0.3246	0.0153	0.1848	0.0185
	VIM-MDI	0.6237	0.0146	0.5164	0.0115	0.4501	0.0107
	VIM-MDII	0.6303	0.0122	0.5164	0.0115	0.4501	0.0107
	VIM-MDIII	**0.6803**	0.0127	**0.6156**	0.0169	**0.6013**	0.0215
leukemia	VIM	0.6004	0.017	0.4473	0.0171	0.3286	0.0183
	VIM-MDI	0.7325	0.011	0.6502	0.0116	0.608	0.0111
	VIM-MDII	0.7326	0.0093	0.65	0.0116	0.6078	0.0111
	VIM-MDIII	**0.7724**	0.0115	**0.7152**	0.0148	**0.6983**	0.0219
p-gene	VIM	0.6935	0.0485	0.7488	0.0469	0.7802	0.043
	VIM-MDI	**0.8047**	0.0433	**0.8335**	0.0411	**0.8146**	0.0421
	VIM-MDII	0.7929	0.0437	0.8327	0.0417	0.8122	0.0427
	VIM-MDIII	0.7566	0.045	0.8033	0.0504	0.7935	0.0393
spect	VIM	0.7389	0.0551	0.7843	0.0577	0.8246	0.0388
	VIM-MDI	0.8093	0.0444	0.8487	0.0362	**0.8625**	0.0315
	VIM-MDII	**0.8193**	0.0401	**0.8488**	0.0373	0.8606	0.0311
	VIM-MDIII	0.7922	0.0512	0.8212	0.0426	0.8447	0.0411
hearts	VIM	0.7898	0.0446	0.8375	0.0467	0.8717	0.0332
	VIM-MDI	**0.9062**	0.0219	0.906	0.0247	0.9267	0.0225
	VIM-MDII	0.9059	0.0224	**0.9066**	0.0243	**0.9268**	0.0221
	VIM-MDIII	0.9009	0.0216	0.9065	0.022	0.928	0.0195

respectively with $ntree = 500, 1000, 2000$. TABLE V shows that the Spearman's Rank Correlation Coefficient of the feature rankings from VIM-MDs are larger than that of VIM on both high-dimensional datasets and low-dimensional datasets, which indicates that as a feature ranking method, the global stability of VIM-MDs is better than VIM. Specifically, the stability of VIM-MDI and VIM-MDII are similar to each other, while on the high-dimensional data

Table VI
RESULTS OF JACCARD INDEX, KUNCHEVA INDEX AND ACCURACY ON THE GIVEN NUMBER OF TOP RANKED FEATURES. K IS THE SIZE OF FEATURE SUBSET (NTREE=2000).

Datasets	k	measure	Jaccard	Kunchev	precision	Dataset	k	measure	Jaccard	Kunchev	precision
Colon	k=50	VIM	**0.4757**	**0.634**	0.8576	p-gene	k=5	VIM	**0.7101**	**0.8051**	0.8475
		VIM-MDI	0.4658	0.6248	0.8552			VIM-MDI	0.6254	0.7369	**0.8641**
		VIM-MDII	0.4622	0.6212	**0.8762**			VIM-MDII	0.6077	0.7174	0.8541
		VIM-MDIII	0.3498	0.5036	0.8648			VIM-MDIII	0.5339	0.6541	0.8524
	k=100	VIM	0.4225	0.5719	0.8731		k=10	VIM	**0.6369**	**0.7251**	**0.9081**
		VIM-MDI	0.4654	0.6147	0.8712			VIM-MDI	0.6116	0.7009	0.9005
		VIM-MDII	**0.4662**	**0.6154**	0.8619			VIM-MDII	0.6264	0.7143	0.903
		VIM-MDIII	0.3387	0.4786	**0.8733**			VIM-MDIII	0.6004	0.6901	0.8999
	k=200	VIM	0.4204	0.5462	0.8681		k=20	VIM	0.6925	0.7176	0.8805
		VIM-MDI	**0.4688**	**0.5974**	0.8564			VIM-MDI	**0.7564**	**0.7843**	**0.9245**
		VIM-MDII	0.4651	0.5937	**0.885**			VIM-MDII	0.7564	0.7843	0.9143
		VIM-MDIII	0.3115	0.4159	0.8802			VIM-MDIII	0.7426	0.7706	0.9007
	k=300	VIM	0.3955	0.49	0.8531		k=30	VIM	0.6873	0.6059	**0.9094**
		VIM-MDI	**0.4818**	**0.588**	**0.8898**			VIM-MDI	0.6908	0.6106	0.9017
		VIM-MDII	0.4803	0.5865	0.86			VIM-MDII	**0.6959**	**0.6184**	0.9025
		VIM-MDIII	0.3301	0.4069	0.865			VIM-MDIII	0.6576	0.5606	0.8974
	k=400	VIM	0.3773	0.4346	0.8448		k=40	VIM	0.6995	0.4039	0.8588
		VIM-MDI	**0.4878**	**0.5693**	0.8295			VIM-MDI	0.7287	0.4691	**0.8949**
		VIM-MDII	0.4869	0.5682	**0.8679**			VIM-MDII	**0.7329**	**0.4784**	0.8871
		VIM-MDIII	0.347	0.3934	0.8555			VIM-MDIII	0.7267	0.4654	0.8697
Leukemia	k=50	VIM	**0.3382**	**0.4969**	0.9245	SPECT	k=5	VIM	**0.6656**	**0.7593**	**0.8125**
		VIM-MDI	0.3266	0.4834	**0.9302**			VIM-MDI	0.655	0.7493	0.7863
		VIM-MDII	0.3337	0.492	0.9243			VIM-MDII	0.6312	0.7242	0.795
		VIM-MDIII	0.1831	0.2999	0.923			VIM-MDIII	0.5392	0.649	0.7838
	k=100	VIM	0.3386	0.4917	0.9075		k=10	VIM	0.5102	0.5715	0.815
		VIM-MDI	0.3417	0.4962	0.9134			VIM-MDI	0.6211	0.6923	0.825
		VIM-MDII	**0.3443**	**0.4987**	**0.9248**			VIM-MDII	0.6211	0.6923	0.8162
		VIM-MDIII	0.1903	0.3047	0.9221			VIM-MDIII	**0.6575**	**0.7268**	**0.8562**
	k=200	VIM	0.2974	0.4346	0.8989		k=20	VIM	**0.6934**	**0.6639**	**0.8325**
		VIM-MDI	**0.3485**	**0.4956**	0.9			VIM-MDI	0.6724	0.6354	0.8075
		VIM-MDII	0.3464	0.4932	**0.912**			VIM-MDII	0.6707	0.6333	0.79
		VIM-MDIII	0.1927	0.2944	0.9004			VIM-MDIII	0.678	0.6415	0.8263
	k=300	VIM	0.2756	0.3957	0.8961		k=30	VIM	**0.8177**	**0.6811**	**0.8175**
		VIM-MDI	0.3472	0.4841	**0.8964**			VIM-MDI	0.8163	0.6787	0.7775
		VIM-MDII	**0.348**	**0.4851**	0.8816			VIM-MDII	0.8126	0.6717	0.7925
		VIM-MDIII	0.195	0.2837	0.8921			VIM-MDIII	0.8058	0.6601	0.7962
	k=400	VIM	0.2579	0.3594	**0.9084**		k=40	VIM	0.8709	0.2361	**0.805**
		VIM-MDI	**0.346**	**0.4721**	0.8964			VIM-MDI	**0.9167**	**0.5172**	0.7712
		VIM-MDII	0.3443	0.4702	0.8598			VIM-MDII	**0.9167**	**0.5172**	0.7838
		VIM-MDIII	0.2043	0.283	0.8661			VIM-MDIII	0.9043	0.4439	0.7875
Central	k=50	VIM	0.1688	0.2821	0.81	hearts	k=5	VIM	0.4974	0.6089	0.783
		VIM-MDI	**0.1964**	**0.3228**	0.815			VIM-MDI	**0.7323**	**0.8145**	**0.7856**
		VIM-MDII	0.1945	0.3201	**0.8217**			VIM-MDII	**0.7323**	**0.8145**	0.7838
		VIM-MDIII	0.1071	0.1872	0.7667			VIM-MDIII	0.6899	0.7844	0.7895
	k=100	VIM	0.1512	0.2517	0.8083		k=10	VIM	**0.6788**	**0.7441**	0.8141
		VIM-MDI	**0.1901**	**0.3092**	**0.83**			VIM-MDI	0.6786	0.7469	0.8211
		VIM-MDII	0.1899	0.309	0.7967			VIM-MDII	0.6786	0.7469	**0.8229**
		VIM-MDIII	0.1092	0.185	0.775			VIM-MDIII	0.6773	0.7441	0.8057
	k=200	VIM	0.1446	0.2506	**0.7967**		k=20	VIM	0.7434	0.725	0.8286
		VIM-MDI	**0.2035**	**0.3185**	0.7867			VIM-MDI	0.7983	0.7922	0.8178
		VIM-MDII	0.2025	0.3171	0.7567			VIM-MDII	0.7983	0.7922	**0.8322**
		VIM-MDIII	0.1216	0.1939	0.79			VIM-MDIII	**0.8161**	**0.8106**	0.8262
	k=300	VIM	0.145	0.2202	**0.785**		k=30	VIM	0.7731	0.5949	0.821
		VIM-MDI	0.2161	0.3268	0.7583			VIM-MDI	**0.8674**	**0.7742**	0.814
		VIM-MDII	**0.2166**	**0.3275**	0.7767			VIM-MDII	0.8646	0.7695	**0.8212**
		VIM-MDIII	0.1311	0.1976	0.775			VIM-MDIII	0.8395	0.723	0.8124
	k=400	VIM	0.1397	0.2001	0.7567		k=40	VIM	0.8883	0.3461	0.8113
		VIM-MDI	0.225	0.3295	0.7533			VIM-MDI	0.9391	0.6517	**0.817**
		VIM-MDII	**0.2258**	**0.3305**	**0.7883**			VIM-MDII	0.9391	0.6517	0.809
		VIM-MDIII	0.1411	0.2024	0.7433			VIM-MDIII	**0.9504**	**0.7189**	0.807

VIM-MDIII achieves the best performances.

TABLE VI shows the results of Jaccard, Kuncheva index and predictive accuracy with different sizes of feature subsets selected by the four methods with the parameter $ntree = 2000$. As the size of selected feature subsets increases, the tendency of Jaccard index is highly consistent with that of Kuncheva index. More specifically, if Jaccard index of VIM-MDs is larger or smaller than that of VIM, then Kuncheva index has the same performances. When k is relatively small, for example, 5 or 50, both the Jaccard and Kuncheva index of VIM-MDs are smaller than VIM, which implies that the stability of the top-ranking features selected by VIM-MDs is undesirable compared to VIM. However, with the value of k increase, VIM-MDs gradually outperform VIM. In particular, the highest predictive accuracies on all the datasets are achieved with feature subsets selected by VIM-MDs which demonstrates their effectiveness.

V. CONCLUSIONS

We propose new variable importance measures for random forest based on margin distribution and in the experiments we investigate the stability and predictive accuracy of both VIM-MDs and VIM. The result shows that the global stability of VIM-MDs is superior to VIM, while they outperform VIM in terms of predictive accuracy. VIM-MDs can be

an effective screening tool for many applications especially for high-dimensional data. We will conduct extensive and comprehensive investigations on our method in the future.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. This work is supported by the State Scholarship Fund of China Scholarship Council under Grant No.201506315008 and Natural Science Foundation of China under Grant No.61673324 and No.61672441, and the Fundamental Research Funds for the Central Universities of China under Grant No.20720160085.

REFERENCES

[1] Kathryn L Lunetta, L Brooke Hayward, Jonathan Segal, and Paul Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1):32, 2004.

[2] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[3] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

[4] Sophia SF Lee, Lei Sun, Rafal Kustra, and Shelley B Bull. Em-random forest and new measures of variable importance for multi-locus quantitative trait linkage analysis. *Bioinformatics*, 24(14):1603–1610, 2008.

[5] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[6] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

[7] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

[8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] M Luz Calle and Víctor Urrea. Letter to the editor: stability of random forest importance measures. *Briefings in bioinformatics*, 12(1):86–89, 2010.

[10] Kristin K Nicodemus. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, 12(4):369–373, 2011.

[11] Huazhen Wang, Fan Yang, and Zhiyuan Luo. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC bioinformatics*, 17(1):60, 2016.

[12] Yang Fan, Li Xuan, Zhou Qifeng, and Luo Linkai. Margin based variable importance for random forest. In *Computer Science & Education (ICCSE), 2011 6th International Conference on*, pages 1361–1366. IEEE, 2011.

[13] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

[14] Amit Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(24):35–43, 2001.