Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 4, April 2019, pp. 1171-1180
DOI: 10.23940/ijpe.19.04.p12.11711180

Similarity based on the Importance of Common Features in Random Forest

Xiao Chen[a,b,*], Li Han[a], Meng Leng[a], and Xiao Pan[c]

[a]*Network Technology Center, Hebei Normal University of Science and Technology, Qinhuangdao, 066004, China*
[b]*Qianan College, North China University of Science and Technology, Qianan, 064400, China*
[c]*College of Economic and Management, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China*

Abstract

In the existing methods for calculating the similarity between samples in random forests, the only case considered is where different samples fall on the same leaf node of the decision tree. The cases where there are leaf nodes in different positions of the decision tree or the sample falls on different leaves are neglected, thus affecting the accuracy of the similarity. In this paper, firstly, according to the difference of the leaf nodes in different positions of the decision tree, the importance of the sample features to which the leaf nodes belong are used as an attribute to describe the similarity. Secondly, for the case that the samples fall on different leaf nodes, the common features between samples are taken as another attribute to describe the similarity. Therefore, the measure method SICF (similarity between samples based on the importance of common features) is proposed. Finally, it is applied to the K-nearest neighbor classification algorithm, and the validity and correctness of the similarity are verified by the OOB index. The experimental results show that for the UCI data set, compared with two classical methods, the similarity SICF achieves better classification results.

Keywords: random forest; similarity between samples; sample feature; feature importance; k-nearest neighbor; classification

(Submitted on November 16, 2018; Revised on December 12, 2018; Accepted on January 6, 2019)

1. Introduction

As one of the branches of data mining, classification methods have been widely used in many fields. According to the number of classifiers, the classification methods can be further divided into single classifier algorithms and multi-classifier algorithms. Among them, decision tree [1], support vector machine (SVM) [2], and naive Bayes (NB) [3] are typical algorithms based on single classifier. To address the performance limitations of single classifiers, multi-classifiers were developed. In multi-classifiers, the data is classified with each meta-classifier and then combined with the classification results according to a certain strategy to obtain the final classification result. Random forest (RF) is a typical multi-classifier algorithm that was proposed by Professor Breiman of the American Academy of Sciences in 2001 [4]. It uses the classification and regression tree (CART) in the decision tree as the meta-classifier. Random forests have high classification performance, fewer parameters to adjust, fast and efficient calculations, and strong noise tolerance. There is no need to worry about over-fitting, and it has been widely used in various industries and fields, such as biological science, finance, chemical engineering, and agricultural science. It has achieved great success and attracted widespread attention.

However, in the existing measure methods of similarity between samples in random forests, the similarity is only associated with the frequency that the different samples appear on the same leaf node of the decision tree. If and only if two samples fall on the same leaf node of the decision tree, the corresponding position of the similarity matrix is increased by 1; otherwise, it is not increased. This method ignores the influence of the position of leaf nodes in the decision tree and the number of common features between leaf nodes on the similarity. Therefore, two aspects are discussed in the following.

A decision tree in a random forest is shown in Figure 1. The decision tree consists of a root node, multiple branch nodes

* Corresponding author.
E-mail address: chenxiao0604@163.com

(sample features), and multiple leaf nodes (sample categories). In Figure 1, the features set of samples is $\{lw, ld, rw, rd\}$ $\{\mathrm{lw}, \mathrm{ld}, \mathrm{rw}, \mathrm{rd}\}$, and the importance of each feature is shown in Table 1.

(1) Assume that both samples x_1 and x_2 fall on the leaf node 1, and both samples x_3 and x_4 fall on the leaf node 3. According to the existing method, their similarity values are all increased by 1. However, leaf nodes 1 and 3 belong to the features *rw*rw and *ld*ld respectively, and the importance of each feature is different on the decision tree. Therefore, it is not reasonable for samples x_1 and x_2 and samples x_3 and x_4 to have the same similarity.

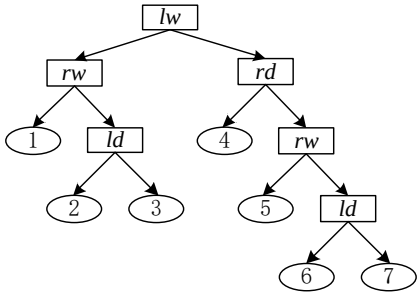


Figure 1. Decision tree *t*

(2) Assume that the samples x_1 and x_2 fall on the leaf nodes 1 and 3 respectively, and the samples x_3 and x_4 fall on the leaf nodes 5 and 7 respectively, and then the similarities are all 0 according to the existing method. However, leaf nodes 1 and 3 have two common features, and leaf nodes 5 and 7 have three common features. Based on the idea that the more common features there are between samples, the greater the similarity, it is unreasonable that the similarity of samples x_1 and x_2 and samples x_3 and x_4 are both 0 in the decision tree.

Table 1. The importance of features

	lw	*ld*	*rw*	*rd*
Imp	0.049	0.045	0.058	0.047

We can see from the above two cases that when the samples fall on the same leaf node of the decision tree, the similarity is uniformly increased by 1, or when they fall on different leaf nodes, the similarity is uniformly increased by 0. Thus the "one size fits all" measure method is unreasonable, and it cannot truly reflect the similarity between samples. In order to improve the accuracy of the similarity, it is necessary to propose a new method for calculating the similarity.

Therefore, firstly, based on the idea that the more important the sample features are, the greater the similarity, the importance of features is regarded as an attribute of describing the similarity. Secondly, based on the idea that the more common features there are between samples, the greater the similarity, the common features are regarded as another attribute. Then, the measure method of similarity between samples based on the common features and its importance is proposed. Finally, the correctness and validity of the similarity measure method are verified by experiments.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 introduces the preliminary, including random forest, OOB estimate, similarity between samples, and feature importance. The similarity based on the importance of common features and the algorithm SICF are proposed formally in Section 4. Section 5 evaluates the validity and effectiveness of the algorithm by experiments. The last section concludes the paper and forecasts the future work.

2. Related Work

Random forest is one of the classical data mining methods. As a late model of the combined classifier algorithm, the idea of random subspace was first proposed by Tin Kam Ho of the British Bell Lab in 1998 [5]; the subspace was randomly selected in the feature space, and then the idea was used by the American Academy of Sciences academician Leo Breiman, who combined bagging integrated learning with the CART decision tree algorithm to propose a random forest algorithm. The random forest algorithm has excellent performance, making the algorithm widely used in various fields, including bioinformatics, ecological environment, economic management, and urban economics.

An important advantage of the random forest algorithm is that it can be used to obtain the sample similarity matrix, and the matrix has a very wide range of applications in all aspects. Pedestrian recognition is one of the most active research directions in the field of pattern recognition. With the enlargement of the sample database, the speed and accuracy of image retrieval and recognition are greatly affected. Wang synthesized various feature extraction algorithms, adopted multi-feature fusion, and proposed a method based on random forest and RankSVM [6]. This optimized pedestrian recognition method uses the potential connection between features of the same pedestrian and the clustering algorithm to explore the potential links between the integrated features clustering the samples. Random forest is used to predict the classification number of the test target, so that the positive sample and the test target are classified into the same category; later, the RankSVM is used to rank the similarity only in the secondary class. Qiu [7] used the random forest sample similarity matrix to map the features and used this sample similarity matrix to calculate the coordinates of each sample. Zhou [8] used the similarity matrix to calculate the similarity between samples in the training set and among samples within classes and then used the eigenvalue random replacement technique to measure the similarity as the characteristic importance evaluation index, thereby sorting all the sample features. Wang [9] obtained the sample similarity matrix based on the random forest model. Based on the similarity matrix, the k-nearest neighbor algorithm was used to classify the data to be detected to determine whether an intrusion occurred.

3. Preliminary

3.1. Random Forest

Random forest [10-12] uses the classification regression tree (CART) as a meta-classifier, combined with bagging's self-sampling method to generate several different training sample sets. In the process of constructing a single tree, a part of the attributes set is randomly selected as the features of the current node, and then optimal features are selected to partition the subtree.

A random forest is usually defined as a set of tree classifiers: $\{h(x,\beta_t), t=1,2,\cdots,ntree\}$ $\{h(x,\beta_t), t=1,2,\ldots,ntree\}$. Among them, the meta-classifier $h(x,\beta_t)$ $h(x,\beta_t)$ is an unpruned classification regression tree generated by the CART algorithm; x is an input vector; and β_t β_t is an independent and identically distributed random vector that determines the generation process of a single tree.

Given a sample set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$, the random forest model $h(x,\beta_t)$ $h(x,\beta_t)$ of $ntree$ decision trees is obtained by training. The specific process is as follows. (1) Suppose that the size of the forest to be built is $t(1\le t\le ntree)$ $t(1\le t\le ntree)$. In the training samples set, the autonomous sampling set D_t D_t is generated by a self-supporting sampling method/bagging method. For any sample set D_t, a classification tree is constructed, and then t classification trees are constructed. (2) For any sample set D_t, the number of sample features is M. The process of constructing a single tree is as follows: for any internal node in the tree, firstly, $m(m<<M)$ $m(m\ll M)$ features are selected randomly from M features as candidate features, and then the optimal features subset is selected by a certain strategy from m features, and finally the subtree is parted by the optimal features. (3) The t trees classifier is constructed in turn to achieve the prediction of unknown samples. (4) In the task of classification, the classification result of unknown sample is the category that is calculated by the maximum number of votes in the $ntree$ tree classifier.

Due to the excellent performance of the random forest algorithm, it has been widely studied and applied in classification, regression, and abnormal point detection.

3.2. OOB Estimate

For each self-service sampling, the original data set (all the samples) is divided into two parts: the training set and the test set, in which the training set accounts for about 63% of the original data set and the test set accounts for about 37%. The 37% of the samples data is also called Out-Of-Bag (OOB). After the classifier is constructed by the training set, the generalization performance of the random forest can be estimated by using the sample of the test set, which is also called the OOB estimation.

For each sample in OOB data, firstly, the classification results of the sample are statistics in the decision tree, which are usually determined by the simple maximum voting method; then, the proportion of the number of samples wrongly

classified in the total number of samples is statistics, that is, the OOB error rate in the random forest. The OOB error rate has been proven by Breiman as an unbiased estimate of the random forest generalization error. Therefore, there is no need to carry out cross-validation or a separate test set to obtain the unbiased estimate of the test set error.

3.3. Similarity Between Samples

The similarity matrix can be obtained by calculating the similarity between samples in the random forest. In existing methods, the similarity is usually measured by the frequency that two samples appear on the same leaf node in each decision tree [13-14]. That is, the higher the frequency, the higher the similarity.

Given a data set $D=\{x_0, x_1, \cdots, x_i, \cdots, x_j, \cdots, x_{N-1}\}$ ($|D|=N$) $\mathrm{D}=\left\{\mathrm{x_0}, \mathrm{x_1}, \ldots, \mathrm{x_i}, \ldots, \mathrm{x_j}, \ldots, \mathrm{x_{N-1}}\right\}(|\mathrm{D}|=\mathrm{N})$, the measure method of the similarity between samples is as follows. (1) Initialize the similarity matrix $Prox$, $Prox_{N\times N}=\{0\}$ $\mathrm{Prox_{N\times N}}=\{0\}$, which is an all-zero matrix of N rows and N columns. (2) For any two samples x_i and x_j, the classification of the unknown sample is predicted by each generated decision tree. If they all fall on the same leaf node of the tree, then 1 is added to the i^{th} row and j^{th} column of the matrix, that is, $Prox_{ij}=Prox_{ij}+1$ $\mathrm{Prox_{ij}=Prox_{ij}+1}$; otherwise, nothing is added. (3) The above process is repeated for each tree in the forest, and then the total appears number $Prox_{ij}$ is obtained. (4) $Prox_{ij}$ will be divided by the size of the random forest, that is, $Prox(x_i, x_j)=Prox(x_i, x_j)/ntree$, $\mathrm{Prox(x_i, x_j)=Prox(x_i, x_j)/ntree}$ and the value is the similarity between samples x_i and x_j. (5) The similarity of all samples is obtained by repeating the above process, and then the similarity matrix $Prox$ is obtained.

It can be seen that the matrix $Prox$ is a symmetric matrix of N rows and N columns; the main diagonal elements are all 1, that is, $Prox(x_i, x_i)=1$ $\mathrm{Prox(x_i, x_i)=1}$ and $Prox(x_i, x_j)=Prox(x_j, x_i)$ $\mathrm{Prox(x_i, x_j)=Prox(x_j, x_i)}$. The similarity $Prox(x_i, x_j)\in[0,1]$ $\mathrm{Prox(x_i, x_j)\in[0,1]}$ between any two samples x_i and x_j. If $Prox(x_i, x_j)$ is closer to 1, the similarity is larger. If it is closer to 0, the similarity is smaller.

3.4. Feature Importance

In the random forest model, the feature importance of sample [15-16] refers to the importance of features in the data set to the category of the predicted sample. The importance of feature is the basis for feature selection.

Given a sample set $D=\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, $\mathrm{D}=\{(\mathrm{x_1}, \mathrm{y_1}), (\mathrm{x_2}, \mathrm{y_2}), \ldots, (\mathrm{x_m}, \mathrm{y_m})\}$ the random forest model $H(x)=\{h(x)_1, h(x)_2, \cdots, h(x)_{ntree}\}$ $\mathrm{H(x)}=\{\mathrm{h(x)_1}, \mathrm{h(x)_2}, \ldots, \mathrm{h(x)_{ntree}}\}$ is obtained by training. Then, $imp(f_k)$ $\mathrm{imp(f_k)}$ is the importance of feature f_k $\mathrm{f_k}$, and its specific calculation process is as follows. (1) Given the decision tree $h(x)_t$ $\mathrm{h(x)_t}$, the prediction error $erra_t$ $\mathrm{erra_t}$ of the decision tree is obtained by calculating the test data set OOB. (2) For all the samples in OOB, the value of feature f_k is changed randomly, and the prediction error $errb_t(f_k)$ $\mathrm{errb_t(f_k)}$ is obtained by calculating again. (3) By repeating (1) and (2), the prediction error of the $ntree$ decision tree before and after the test data change can be obtained. Then, the importance of the sample feature f_k, that is, $imp(f_k)$, is shown in Equation (1).

$$imp(f_k)=\frac{1}{ntree}\sum_{t=1}^{ntree} errb_t(f_k)-erra_t \tag{1}$$

In Equation (1), the importance of feature is evaluated by using the average value of prediction error difference of $ntree$ decision trees before and after the change of test data. The more important the feature, the greater the prediction error, and then the variation of the prediction error difference is also greater. Therefore, Equation (1) is a reasonable method for assessing the importance of features.

4. Similarity based on the Importance of Common Features

4.1. Similarity Between Samples

In the existing methods of similarity in random forests, some problems are ignored, such as the different position of leaf nodes in the decision tree and the different numbers of common features. Therefore, a new method is proposed for calculating similarity between samples in this paper. For the convenience of description, the relevant definitions are as follows.

Definition 1 (Sample Path) Given a decision tree $t(1 \le t \le ntree)$ $\mathrm{t}(1 \le \mathrm{t} \le \mathrm{ntree})$, for any sample x_i, the node sequence from the leaf node (to which it belongs) to the root node constitutes the path $L(x_i)_t$ $\mathrm{L}(\mathrm{x_i})_\mathrm{t}$ of the sample x_i on the decision tree. The length of the path is recorded as $l(x_i)_t$ $\mathrm{l}(\mathrm{x_i})_\mathrm{t}$.

Definition 2 (Sample Features Set) Given a decision tree t, for any sample x_i $\mathrm{x_i}$, its features set is the branch nodes included in the path of the decision tree t, denoted as $F(x_i)_t = \{f_k \mid f \in L(x_i)_t\}$. $\mathrm{F}(\mathrm{x_i})_\mathrm{t} = \{\mathrm{f_k} | \mathrm{f_k} \in \mathrm{L}(\mathrm{x_i})_\mathrm{t}\}$

Definition 3 (Common Features Set) Given a decision tree t, for any samples x_i and x_j, the common features set between samples is recorded as $CF(x_i, x_j)_t$ $\mathrm{CF}(\mathrm{x_i}, \mathrm{x_j})_\mathrm{t}$, as shown in Equation (2). The common features number is recorded as $|CF(x_i, x_j)_t|$ $\left|\mathrm{CF}(\mathrm{x_i}, \mathrm{x_j})_\mathrm{t}\right|$.

$$CF(x_i, x_j)_t = F(x_i)_t \cap F(x_j)_t \tag{2}$$

For example, in the decision tree, as shown in Figure 1, when the sample x_1 $\mathrm{x_1}$ falls on the leaf node 1, its path is $L(x_1)_t = \{lw, rw, x_1\}$ $\mathrm{L}(\mathrm{x_1})_\mathrm{t} = \{\mathrm{lw}, \mathrm{rw}, \mathrm{x_1}\}$, and the path length $l(x_1)_t$ $\mathrm{l}(\mathrm{x_1})_\mathrm{t}$ is 2. The features set of samples x_1 is $F(x_1)_t = \{lw, rw\}$ $\mathrm{F}(\mathrm{x_1})_\mathrm{t} = \{\mathrm{lw}, \mathrm{rw}\}$.

Considering the correlation between similarity and the position of the leaf nodes on the decision tree (that is, the importance of features), when two samples fall on the same leaf node of the decision tree, the sum of the importance of all the features from the leaf node to the root node is taken as the similarity between samples. For example, when the samples x_1 and x_2 $\mathrm{x_2}$ fall on the leaf node 1, then the sum of the feature importance is 0.049 + 0.058 = 0.107. Similarly, when both samples x_3 $\mathrm{x_3}$ and x_4 $\mathrm{x_4}$ fall on the leaf node 3, the features set is $F(x_3)_t = F(x_4)_t = \{lw, rw, ld\}$ $\mathrm{F}(\mathrm{x_3})_\mathrm{t} = \mathrm{F}(\mathrm{x_4})_\mathrm{t} = \{\mathrm{lw}, \mathrm{rw}, \mathrm{ld}\}$, and the sum of the feature importance is 0.049 + 0.058 + 0.045 = 0.152. Considering again that different decision trees select different features, the importance of features is regularized, and a new method for calculating the importance of features in each decision tree is proposed.

Definition 4 (Feature Importance) Given the decision tree t, the importance of feature f_k is recorded as $imp(f_k)_t$ $\mathrm{imp}(\mathrm{f_k})_\mathrm{t}$, as shown in Equation (3).

$$imp(f_k)_t = \frac{imp(f_k)}{total_t} \tag{3}$$

Among them, $total_t$ $\mathrm{total}_\mathrm{t}$ is the sum of the importance of all the features in the decision tree t; the calculation method of $imp(f_k)$ is shown in Equation (1).

Based on the idea that the higher the importance of sample features, the greater the similarity, and the more common features there are between samples, the greater the similarity, the measure method for the similarity between samples is given as follows.

Definition 5 (Similarity between Samples) Given the decision tree t, for any samples x_i and x_j, the similarity between samples is recorded as $Prox(x_i, x_j)_t$ $\mathrm{Prox}(\mathrm{x_i}, \mathrm{x_j})_\mathrm{t}$, as shown in Equation (4).

$$Prox(x_i, x_j)_t = \frac{\sum_{f_k \in CF(x_i, x_j)_t} imp(f_k)_t}{Min\left(\sum_{f_k \in CF(x_i)_t} imp(f_k)_t, \sum_{f_k \in CF(x_j)_t} imp(f_k)_t\right)} \tag{4}$$

We can see from Equation (4) that when two samples fall on the same leaf node of the decision tree, the similarity is 1, which is consistent with the original similarity calculation method.

4.2. Algorithm of Similarity Between Samples

Based on the idea that the more common features there are between samples and the more important the features, the greater the similarity, a new measure method of the similarity SICF is proposed in the random forest. The specific algorithm is as follows.

Algorithm 1 SICF (Similarity based on the Importance of Common Features)

Input: samples set *D*, random forest model $H(x) = \{h(x)_1, h(x)_2, \cdots, h(x)_{ntree}\}$ $\mathrm{H(x)} = \{\mathrm{h(x)_1, h(x)_2, \ldots, h(x)_{ntree}}\}$

Output: similarity matrix *Prox*

(1) For $i = 1$ to m
(2) For $j = 1$ to m
(3) For $t = 1$ to *ntree*
(4) $Prox(x_i, x_j) += Prox(x_i, x_j)_t$
(5) End For
(6) $Prox(x_i, x_j) = Prox(x_i, x_j) / ntree$
(7) End For

4.3. K-Nearest Neighbor Classification based on SICF

K-nearest neighbor is one of the classical classification techniques. It does not need to study the training set in advance and obtain the model and only needs to select a suitable similarity index. The Euclidean distance is usually used as a similarity index. The specific classification process of the test samples is as follows. Firstly, calculate the similarity between the test sample and the training sample. Secondly, select K samples that are most similar to it in the training set. Finally, calculate which of the K samples has the highest number of occurrences and if multiple categories have the same maximum frequency; then, one of them is randomly selected, and it is the prediction result. The specific process of K-nearest neighbor classification based on improved similarity is shown in Figure 2.

As can be seen from Figure 2, there are five main steps. (1) The data set *D* is divided into two parts by the hold-out method, which is the training set *S* and the test set *T* respectively, where the ratio of the number of samples contained in *S* and *T* is 7:3. (2) The random forests model $H(x) = \{h(x)_1, h(x)_2, \cdots, h(x)_{ntree}\}$ $\mathrm{H(x)} = \{\mathrm{h(x)_1, h(x)_2, \ldots, h(x)_{ntree}}\}$ is constructed on the training set *S*. (3) In the random forest model $H(x)$, the similarity matrix *Prox* is obtained by calculating the similarity between each sample in test set *T* and in training set *S*. In this matrix, the sample in *T* is the row label and the sample in *S* is the column label. (4) According to the similarity matrix *Prox*, calculate the mean value of similarity between test samples and each category sample. (5) Select the category with the maximum average similarity as the prediction result of the sample to be tested.

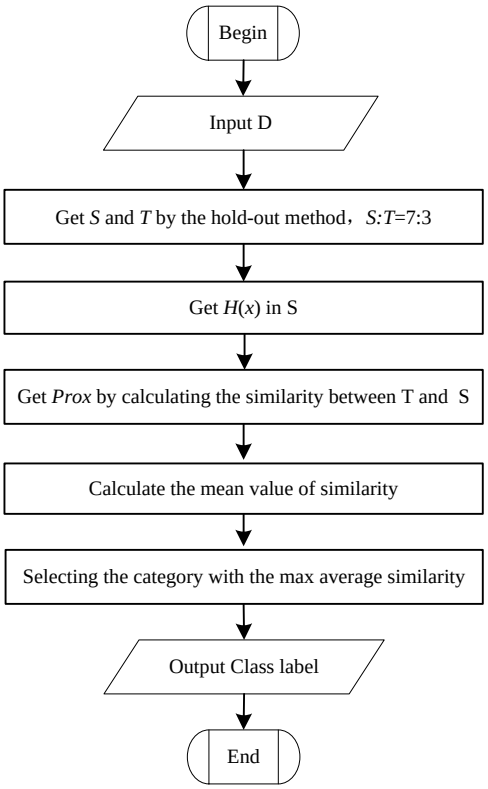


Figure 2. The process of K-nearest neighbour classification based on SICF

5. Experiments

In order to verify the effectiveness of the improved similarity method, it is applied to the K-nearest neighbor classification algorithm, and it is the similarity index of the K-nearest neighbor algorithm. OOB is used as the evaluation index, that is, the ratio of the number of samples with the correct classification to the total number of samples in the test set is calculated in order to evaluate the advantages and disadvantages of the similarity in the random forest.

5.1. Experimental Scheme and Parameter Setting

The experimental data sets consist of ten groups' data sets in the UCI, which are respectively Hayes-roth, Parkinsons, Sonar, Seed, Glass, Ecoli, LibrasMovement, Balance, Diabetes, and Car, as shown in Table 2. In this table, we introduce the basic information of data sets, such as the number of samples, number of features, and number of categories.

Table 2. The basic information of data sets

Data set	Number of samples	Number of features	Number of categories
Hayes-roth	160	3	3
Parkinsons	195	22	2
Sonar	208	60	2
Seed	210	7	3
Glass	214	9	6
Ecoli	336	7	8
LibrasMovement	360	90	15
Balance	625	4	3
Diabetes	768	8	2
Car	1728	6	4

In order to verify the effectiveness of the improved measure method for the similarity between samples, comparison experiments with the original method and the method of Li Zhengui [17] are carried out on the above data sets. In order to prevent random interference during the experiments, the average value of 100 times random experimental results is taken as the final experimental result. Given the size of the features set of the current node is dd, $\sqrt{d}$ features subset is first randomly selected for each feature partition, and then the optimal feature is selected from the features subset for node splitting [18]. In order to study the effects of different random forest sizes on the experimental results, the size of the

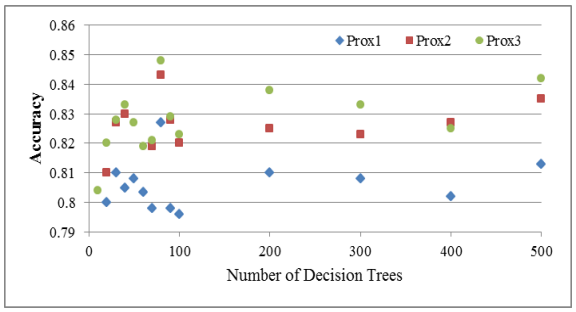
random forest is {10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500}.

5.2. Experimental Results and Analysis

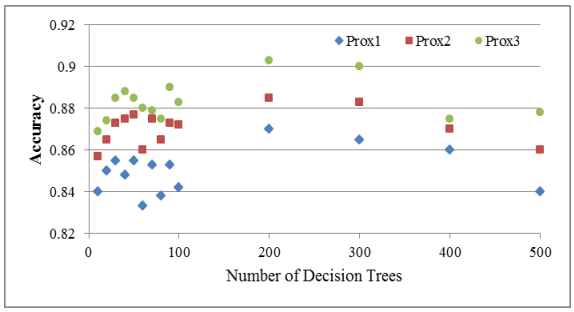
Three methods for calculating the similarity between samples of random forest are applied to the K-nearest neighbor classification algorithm. When the size of decision trees increases from 10 to 500, the experimental results of three indexes based on the classification accuracy are shown in Figure 3. In Figure 3, the ordinate axis represents the average accuracy of the K-nearest neighbor classification of the test sample similarity, and the abscissa axis represents the size of decision trees in the random forest. *Prox*1 represents the original method of similarity between samples, *Prox*2 represents the similarity that is improved by Li Zhengui, and *Prox*3 represents the improved similarity SICF in this paper.

The experimental results on the different data sets are shown in Figure 3. On the Hayes-roth data set, as shown in Figure 3(a), we can see that under the average accuracy, *Prox*2 is 1.9% higher than *Prox*1, and *Prox*3 is 0.2% higher than *Prox*2. On the Parkinsons data set, as shown in Figure 3(b), *Prox*2 is 2.1% higher than *Prox*1, and *Prox*3 is 1.1% higher than *Prox*2. On the Sonar data set, as shown in Figure 3(c), *Prox*2 is 1% higher than *Prox*1, and *Prox*3 is 0.8% higher than *Prox*2. On the Seed data set, as shown in Figure 3(d), *Prox*2 is 0.8% higher than *Prox*1, and *Prox*3 is 0.3% higher than *Prox*2. On the Glass data set, as shown in Figure 3(e), *Prox*2 is 2.2% higher than *Prox*1, and *Prox*3 is 0.8% higher than *Prox*2. On the Ecoli data set, as shown in Figure 3(f), *Prox*2 is 1.6% higher than *Prox*1, and *Prox*3 is 0.7% higher than *Prox*2. On the LibrasMovement data set, as shown in Figure 3(g), *Prox*2 is 1.9% higher than *Prox*1, and *Prox*3 is 1.9% higher than *Prox*2. On the Balance data set, as shown in Figure 3(h), *Prox*2 is 9.1% higher than *Prox*1, and *Prox*3 is 0.5% higher than *Prox*2. On the Diabetes data set, as shown in Figure 3(i), the effect of *Prox*2 and *Prox*1 is almost the same, and *Prox*3 is 2.2% higher than *Prox*2. On the Car data set, as shown in Figure 3(j), *Prox*2 is 0.77% higher than *Prox*1, and *Prox*3 is 0.93% higher than *Prox*2.

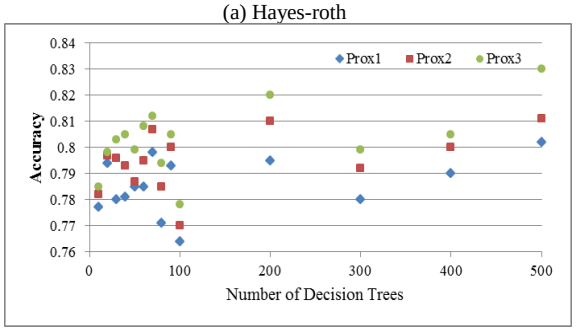
In summary, in the above ten data sets, from the test average accuracy point of view and compared with the other two methods, the classification efficiency based on the similarity with importance of common features is superior. On the Hayes-roth data set, the method based on *Prox*3 compared to *Prox*2 has slightly more accuracy, but *Prox*3's improvement is particularly obvious on the Parkinson, Diabetes, LibrasMovement, and car data sets, which increased by 1.1%, 2.2%, 1.9%, and 0.93% respectively. It can be seen that the similarity between samples can be described more comprehensively by incorporating the common features and the importance of features into it.



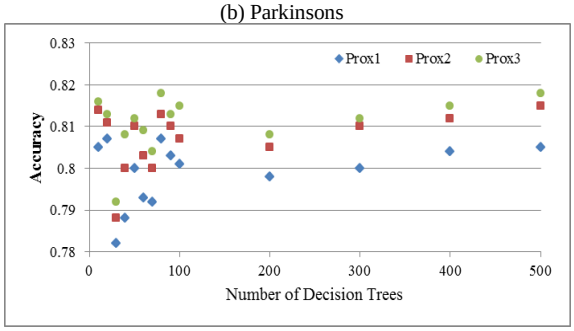
(a) Hayes-roth



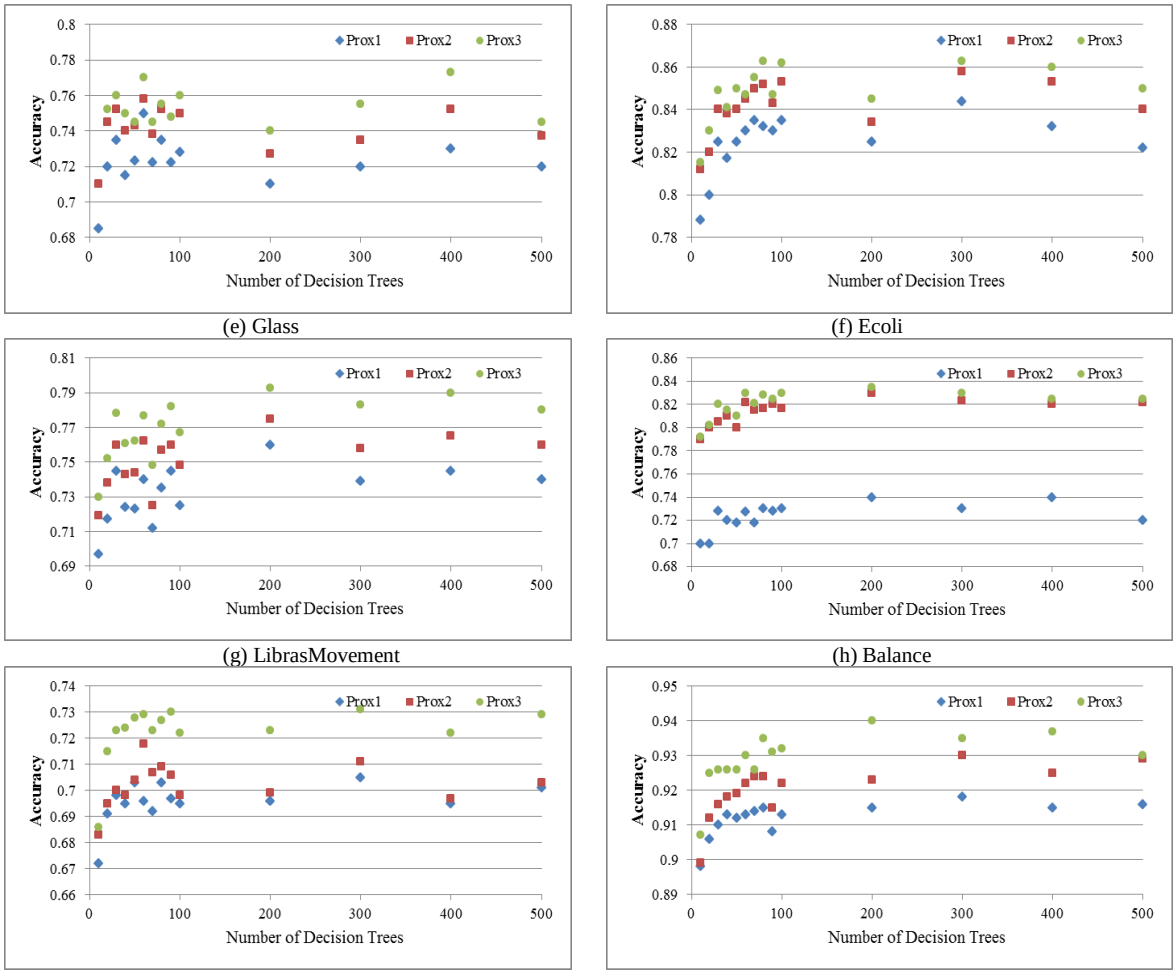
(b) Parkinsons



(c) Sonar



(d) Seeds



(e) Glass (f) Ecoli

(g) LibrasMovement (h) Balance

(i) Diabetes (j) Car

Figure 3. The experimental results on the different data sets

6. Conclusions

In this paper, the common features between samples and the importance of each feature in the decision tree are integrated into the existing similarity, and the measure method of SICF (similarity between samples based on the importance of common features) in random forest is proposed. It is used as a similarity index in K-nearest neighbor classification algorithm, and the correctness and validity of the similarity method have been verified in the UCI data set. The experimental results show that compared with the other two methods, the similarity SICF has a better classification effect, especially in the Parkinsons, Diabetes, LibrasMovement, and Car data sets, which are improved by 1.1%, 2.2%, 1.9%, and 0.93% respectively. On the basis of ensuring further improvement of the accuracy of similarity, improving the execution efficiency of the algorithm is the focus of our next step.

Acknowledgements

This work is supported by the National Youth Science Foundation of Hebei (No. F2017209070), the National Science Foundation of China, (No. 61472340, 61303017), the National Youth Science Foundation of China (No. 61602401), and the Natural Science Foundation of Hebei Province (No. F2014210068).

References

1. S. Shan, "Decision Tree Learning," New York: Springer US, pp. 1-28, February 2016
2. A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine," New York: Springer US, pp. 24-52, 2016
3. K. Adi, C. E. Widodo, A. P. Widodo, et al., "Naïve Bayes Algorithm for Lung Cancer Diagnosis using Image Processing Techniques," *Advanced Science Letters*, Vol. 23, No. 3, pp. 2296-2298, March 2017
4. L. Breiman, "Random Forest," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, January 2001
5. T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis &*

Machine Intelligence, Vol. 20, No. 8, pp. 832-844, August 1998
6. D. Wang, Y. L. Chen, X. D. Cai, et al., "Person Re-Identification based on Random Forest and RankSVM Optimization," *Video Engineering*, Vol. 39, No. 18, pp. 90-93, September 2015
7. Y. H. Qiu, "Customer Loss Prediction in Telecom Industry based on Pruning Random Forest," *Journal of Xiamen University* (*Natural Science Edition*), Vol. 53, No. 6, pp. 817-823, June 2014
8. Q. F. Zhou, W. C. Hong, and F. Yang, "Feature Selection based on Difference Random Forest Similarity Matrix," *Journal of Huazhong University of Science and Technology* (*Natural Science Edition*), Vol. 38, No. 4, pp. 58-61, April 2010
9. H. Wang and H. Z. Yan, "Similar Performance Intrusion Detection Algorithm based on Random Forest Computing," *Information Security and Communication Secrecy*, Vol. 2009, No. 8, pp. 70-73, August 2009
10. Y. Dong, B. Du, and L. Zhang, "Target Detection based on Random Forest Metric Learning," *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, Vol. 8, No. 4, pp. 1830-1838, April 2017
11. L. Huang, Y. Jin, and Y. Gao, "Longitudinal Clinical Score Prediction in Alzheimer's Disease with Soft Split Sparse Regression based on Random Forest," *Neurobiology of Aging*, Vol. 46, No. 10, pp. 180-183, October 2016
12. S. S. Matin and S. C. Chelgani, "Estimation of Coal Gross Calorific Value based on Various Analyses by Random Forest Method," *Fuel*, Vol. 177, No. 8, pp. 274-278, August 2016
13. K. R. Gray, P. Aljabar, and R. A. Heckemann, "Random Forest-based Similarity Measures for Multimodal Classification of Alzheimer's Disease," *Neuroimage*, Vol. 65, No. 1, pp. 167-175, January 2013
14. Y. Qi, J. K. Seetharaman, and Z. B. Joseph, "Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple Sources," *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, Vol. 10, pp. 531-542, 2005
15. H. Y. Lu, M. Zhang, and Y. Q. Liu, "Feature Importance Analysis and Enhanced Feature Selection Model of Convolutional Neural Networks," *Journal of Software*, Vol. 28, No. 11, pp. 2879-2890, November 2017
16. D. Zhang, Q. Wang, B. Zhu, et al., "Pedestrian Recognition using the Importance of Human Body Features," *Journal of Wuhan University* (*Information Science Edition*), Vol. 42, No. 1, pp. 84-90, January 2017
17. Z. G. Li, "Several Studies on The Improvement of Random Forest," *Xiamen*: *Master's Thesis of Xiamen University*, pp. 18-27, 2014
18. Y. Y. Chen, J. Q. Wu, and K. J. Xu, "Attribute Splitting Method based on Gini Index in Decision Tree," *Microcomputer Development*, Vol. 14, No. 15, pp. 66-68, July 2004

Xiao Chen graduated from the College of Information Science and Engineering at Yanshan University with a master's degree and Ph.D. She is a lecturer at Hebei Normal University of Science and Technology. Her research interests include graph mining, social network analysis, and machine learning.

Li Han graduated from the College of Information Science and Engineering at Yanshan University with a bachelor's degree, master's degree, and Ph.D. She is a lecturer at Hebei Normal University of Science and Technology. Her research interests include wireless sensor network and machine learning.

Meng Leng graduated from the College of Information Science and Engineering at Yanshan University, China with a master's degree. He is an associate lecturer at Hebei Normal University of Science and Technology. His research interests include wireless sensor network and machine learning.

Xiao Pan is an associate professor at Shijiazhuang Tiedao University. She was a visiting scholar in the Department of Computer Science at the University of Illinois. She received her Ph.D. in computer science from Renmin University of China in 2010. Her research interests include data management on moving objects, location based social networks, and privacy-aware computing.