Front.Comput.Sci.
DOI

RESEARCH ARTICLE

Structural Diversity for Decision Tree Ensemble Learning

Tao SUN, Zhi-Hua ZHOU (✉)

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Abstract Decision trees are a kind of off-the-shelf predictive models, and they have been successfully used as the base learners in ensemble learning. To construct a strong classifier ensemble, the individual classifiers should be accurate and diverse. However, diversity measure remains a mystery although there were many attempts. We conjecture that a deficiency of previous diversity measures lies in the fact that they consider only *behavioral diversity*, i.e., how the classifiers behave when making predictions, neglecting the fact that classifiers may be potentially different even when they make the same predictions. Based on this recognition, in this paper, we advocate to consider *structural diversity* in addition to behavioral diversity, and propose the TMD (tree matching diversity) measure for decision trees. To investigate the usefulness of TMD, we empirically evaluate performances of selective ensemble approaches with decision forests by incorporating different diversity measures. Our results validate that by considering structural and behavioral diversities together, stronger ensembles can be constructed. This may raise a new direction to design better diversity measures and ensemble methods.

Keywords ensemble learning, structural diversity, decision tree

1 Introduction

A decision tree is a predictive model, which recursively partitions the feature space into subspaces that constitute the bases for prediction. Decision trees are a kind of off-the-shelf models and have been widely used in data science, such as data mining, pattern recognition, bioinformatics [1] and finance [2].

The popularity of decision trees is attributed to the many benefits they offer, including good comprehensibility, being able to handle a variety of input forms and process errors and missing values [3]. One of biggest drawbacks is their instability. Small variations in the training data may result in totally different decision trees. However, this disadvantage can be mitigated by utilizing ensemble learning strategies to grow a decision forest.

Ensemble methods [4] train and combine multiple base learners for one single task, and are a kind of powerful machine learning paradigm. It is widely accepted that ensemble methods usually achieve better generalization performance than single learners. The properties of decision trees make them an ideal choice as the base learnes to ensemble learning, and their combination has led to many successful predictive models, such as random forest [5], extremely randomized trees [6] and rotation forest [7].

To achieve a strong classifier ensemble, the component classifiers should be accurate and more importantly diverse [4]. Combining only accurate classifiers is often worse than combining some accurate ones together with some relatively weak ones, since complementarity is more important than pure accuracy [8]. Many effective heuristic mechanisms have been designed to generate diversity, such as manipulating the training data [5, 9], learning parameters [10], and output representations [11]. In particular, the successes of the decision forests mentioned above are largely credited to their delicate mechanisms to generate diversity.

Despite these successes, the concept of ensemble diversity remains a mystery. Currently, there is no generally accepted formal formulation and measures for ensemble diversity. Plenty of diversity measures have been proposed, such as

Received month dd, yyyy; accepted month dd, yyyy

E-mail: *zhouzh@nju.edu.cn*

κ-statistic, Disagreement and Entropy [4]. However, the effectiveness of existing diversity measures are doubtful since there seems to be no clear relation between those diversity measures and the generalization performance of the ensemble [12–14]. Besides, these measures are closely related to the average individual accuracies. This is undesirable since it is not expected that the diversity measure becomes another estimate of accuracy [4].

We conjecture that a deficiency of previous diversity measures lies in the fact that they consider only *behavioral diversity*, i.e., how the classifiers behave when making predictions, neglecting the fact that classifiers may be potentially different even when they make the same predictions. Consider the simple task of predicting if $Y \geq X$, where X and Y are two components of an instance and both lie in $[0, 10]$. Figure 1 plots the decision regions of two decision trees. The two classifiers make the same predictions on the eight instances; however, they use totally different partitions of the feature space and are obviously different. As another example, supposing that there are redundancy in the features, two classifiers can make exactly the same predictions while using different features. Consequently, they may have different capability of handling missing and drifting in the features.

In this paper, we advocate to consider *structural diversity* as a complement to behavioral diversity. The structure of a learner embeds some inherent properties, such as tendency to overfit, thus is quite important. Unlike behavioral diversity, structural diversity does not rely on specific data. Take decision tree as an example: deeper trees make predictions based on a longer sequence of tests and therefore intuitively tend to be more specialized than shallow trees [15] while shallow trees tend to be more general and knowledge-like [16]. Besides, the similarity in local structure of decision trees corresponds to the similarity in their partitions of the feature space. Thus, it would be beneficial to encourage diversity in the structure. Based on these observations, we propose the TMD (tree matching diversity) measure through the minimum number of node operations to match the structure of two decision trees, which is simple to measure the structural diversity and easy to understand. Since the structural diversity measure might not be enough to describe the difference of two classifiers alone, it should be combined with the behavioral diversity measures. To validate its usefulness, we take the selective ensemble, or called ensemble pruning [4], as a platform, wherein it is relatively easy to substitute different diversity measures and compare their effects on the ensemble performance. We choose three types of selective ensemble approaches and incorporate different diversity measures. Our empirical studies validate that by considering structural and behavioral diversities together, stronger ensembles can be constructed. In this paper, we use the TMD measure for its simplicity, yet meritorious improvements are obtained when it is used to enhance the behavioral diversity measures, showing the necessity to consider structural diversity. Better results are expected by making more delicate use of structural diversity. This may raise a new direction to design better diversity measures and ensemble methods.

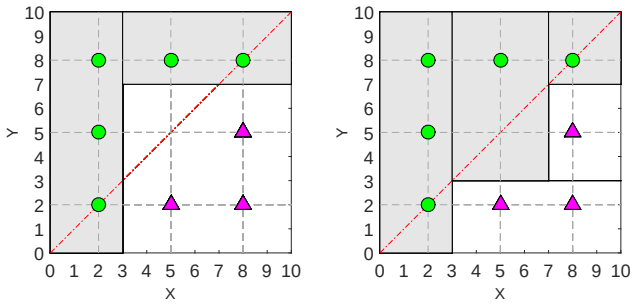


Fig. 1 Decision regions of two decision trees for the task of predicting if $Y \geq X$. Green circles denote positive instances while magenta triangles denote negative instances. Red lines indicate ground-truth decision boundaries. Grey shaded regions are predicted as positive. The two decision trees are obviously different even though making exactly the same predictions on the eight instances.

The rest of this paper is organized as follows: Section 2 reviews previous work on decision tree ensemble learning, diversity measures and selective ensemble methods. The importance of structure is briefly discussed. The definition of tree matching diversity measure and illustrative examples are described in Section 3. Section 4 presents an empirical analysis of the usefulness of tree matching diversity measure under selective ensemble. Conclusions and future work are given in Section 5.

2 Related Work

Decision trees originates from decision theory and statistics, and are enhanced by researchers in other fields like data mining and machine learning. A decision tree can be represented as an acyclic-directed tree model, which consists of a set of internal nodes and decision nodes. Each internal node routes instances to its decendents based on certain criterions and each decision node makes a prediction. One of biggest drawbacks of decision trees is their high variance. A minor change in one split close to the root has a large

disturbance on the whole subtree below. Thus, decision trees are usually used in accompany with ensemble learning to make decision forests. By combining a number of decision trees to make predictions, their variance can be greatly reduced. Meanwhile, the difference among the decision trees is a key to the success of decision forests. Many decision forests use delicate mechanisms to enhance the difference. Random forest randomly samples training data and selects candidate subsets of attributes for splitting [5]. Extremely randomized trees further randomly select both the splitting attributes and its corresponding cut-points [6]. Rotation forest uses PCA to transform training data into several rotated feature spaces [7]. In addition, there are some boosting-like decision forests such as alternating decision tree [17] and gradient boosted decision tree [18].

Ensemble diversity, i.e., the difference among the individual learners, is a fundamental issue in ensemble methods. It is easy to understand that no performance improvement can be obtained if identical learners are combined. However, there is no generally accepted definition of ensemble diversity. The benefit of combining different learners can be seen from error decompositions such as error-ambiguity decomposition [19] and bias-variance-covariance decomposition [20], but these decompositions do not provide a unified formal formulation of ensemble diversity. To measure ensemble diversity, a classical approach is to measure the difference of predictions by two learners and then average all the pairwise measurements.

Given a data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, for binary classification (i.e., $y_k \in \{+1, -1\}$), we have the contingency table for two classifiers h_i and h_j as shown in Table 1, where $a + b + c + d = m$ are non-negative variables showing the numbers of instances satisfying the specific conditions by the corresponding rows and columns.

Table 1 Contingency table for binary classification.

	$h_i = +1$	$h_i = -1$
$h_j = +1$	a	c
$h_j = -1$	b	d

The definitions of three representative diversity measures, κ-statistic (κ), Disagreement (DIS), and Entropy (ENT), are listed in Table 2. Entropy is non-pairwise, in whose definition $P(y|\mathbf{x}_k) = \frac{1}{T}\sum_{i=1}^{T}\mathbb{I}(h_i(\mathbf{x}_k) = y)$ can be estimated by the proportion of individual classifiers that predict y as the label of $\mathbf{x}_k$. $\mathbb{I}(\cdot)$ takes 1 if the statement in the brackets is true and 0 otherwise. For multi-class classification, κ-statistic follows the definition in [21] and Disagreement follows that in [12] by calculating on the oracle outputs. The generalization of Entropy is natural by summing over all the classes.

Table 2 Definitions of three representative diversity measures for binary classification.

Diversity Measure	Definition
κ-statistic	$\kappa_{ij} = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2}$ $\quad \Theta_1 = \frac{a+d}{m}$ $\quad \Theta_2 = \frac{(a+b)(a+c)+(c+d)(b+d)}{m^2}$
Disagreement	$dis_{ij} = \frac{b+c}{m}$
Entropy	$ent = \frac{1}{m}\sum_{k=1}^{m}\sum_{y\in\{+1,-1\}} -P(y|\mathbf{x}_k)\log P(y|\mathbf{x}_k)$

Many other diversity measures are defined in a similar way. Despite that these measures reflect the intuitive notion of diversity, none of them have been derived from an exact decomposition of the ensemble error. Kuncheva and Whitaker [12] presented possibly the first doubt on diversity measures. They did a broad range of experiments on ten diversity measures, showing that there seems to be no clear relation between those diversity measures and the ensemble performance. Tang et al. [13] showed that exploiting the above diversity measures explicitly is ineffective in constructing consistently stronger ensembles compared to algorithms that seek diversity implicitly. The change of existing diversity measures does not provide consistent guidance on whether an ensemble generalizes well. In addition, these measures are closely related to the average individual accuracies, which is undesirable since it is not expected that the diversity measure becomes another estimate of accuracy [4].

Brown and Kuncheva [22] proposed the "good" and "bad" diversity by decomposing the majority voting error for binary classification. They showed that increasing "good" diversity reduces ensemble error while increasing "bad" diversity increases ensemble error. Similar decompositions for multi-class classification are derived in [14]. However, they both depend on the combining rule and are intractable in practice. Information theoretic diversity provides a promising direction for understanding ensemble diversity [23, 24]. The concept of ensemble diversity, however, still remains an open problem. A further investigation and new insights are desired to unveil the mystery.

We conjecture that ensemble diversity is related to the structural diversity in addition to the behavioral diversity, while all previous diversity measures have neglected this. For example, two decision trees may have totally different structure even when they make the same predictions. The structure of a decision tree embeds some inherent properties. Deeper trees make predictions based on a longer sequence of

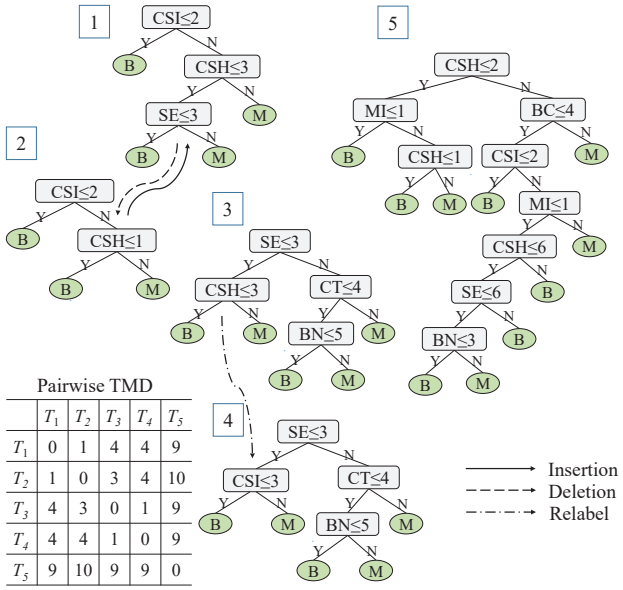
tests and therefore intuitively tend to be more specialized than shallow trees and thus more likely to overfit [15]. The term structure, however, is not so apparent as it sounds. Frederick P. Brooks, Jr. summarized three great challenges for half-century-old computer science in the 50th year of *Journal of the ACM* [25]. The first challenge is *quantification of structural information*. We have no theory that gives a metric for the information embedded in structure. After over a decade, this remains a fundamental issue and needs further insights.

Selective ensemble, also called as ensemble pruning, is a kind of ensemble methods [4]. Instead of combining all the base learners, selective ensemble selects only a subset of them. Obviously, this can improve storage and computational efficiency for predictions. Furthermore, both theoretical and empirical studies have shown that the pruned ensemble can achieve better generalization performance than the whole ensemble [26, 27]. Previous pruning methods can be categorized into three branches, ordering-based, clustering-based and optimization-based. The ordering-based methods commonly start from an empty set and then iteratively add a base learner based on a certain criterion [27]. The clustering-based methods usually employ a clustering process to partition the base learners into a number of groups and then identify a number of representative prototype learners for each group [28, 29]. The optimization-based methods formulate ensemble pruning as an optimization problem, which aims to maximize or minimize an objective related to the generalization performance of the final ensemble [26, 30–32].

3 Tree Matching Diversity

A decision tree consists of a set of nodes and branches. Each non-leaf node is associated with a feature to split and each leaf node is associated with a class label. The structure of a decision tree is quite important. Deep trees and shallow trees may have totally different properties. Besides, the order of nodes matters a lot as instances are passed from root to leaf through nodes and branches sequentially. However, to give an appropriate metric of the structural difference between two decision trees is really hard. Here we propose the *tree matching diversity (TMD)* measure, which is simple to implement and easy to understand. To measure the structural difference, we just neglect the leaf nodes as it is inappropriate to directly compare them to those non-leaf nodes. Future work might consider better utilizing this labeling information.

The pairwise tree matching diversity measure is defined as the minimum number of node operations to match the structure of two decision trees, where three node operations are considered, including inserting a node, deleting a node and replacing the associated feature of a node. This value is no larger than the total node number of the two decision trees, thus is well-defined. It is closely related to the tree edit distance, which can be calculated using strategies like dynamic programming or more efficient decomposition strategies [33]. Note that this measure is symmetric for the two decision trees. The larger the value is, the more different their structures are. The overall diversity of an ensemble with more than two decision trees is defined as the average of all pairwise measures. The measure could be normalized by the maximum pairwise value among the ensemble.



Pairwise TMD

	T_1	T_2	T_3	T_4	T_5
T_1	0	1	4	4	9
T_2	1	0	3	4	10
T_3	4	3	0	1	9
T_4	4	4	1	0	9
T_5	9	10	9	9	0

Fig. 2 Five decision trees. The arrows indicate the node operations needed to match the two trees. Pairwise TMD are listed.

For an illustration, five decision trees are trained on the *Breast Cancer Winsconsin* data set from UCI [34] and plot in Figure 2. The data set has nine attributes, Clump Thickness (CT), Cell Size Uniformity (CSI), Cell Shape Uniformity (CSH), Marginal Adhesion (MA), Single Epithelial Cell Size (SE), Bare Nuclei (BN), Bland Chromatin (BC), Normal Nucleoli (NN) and Mitoses (MI). The task is to predict whether the tumor is Benign (B) or Malignant (M). Pairwise TMD are listed in the figure. The TMD between T_1 and T_2 is 1 as only one node Insertion/Deletion is needed to match the two trees. The TMD between T_3 and T_4 is also 1 as only one node Relabel is needed. All the four trees have a large

TMD with T_5. Hence it is expected that T_1 and T_2 would perform similarly, T_3 and T_4 would perform similarly, while all four would perform differently from T_5. The TMD of the ensemble would be the average of 10 pairwise TMD, i.e., 5.4. Furthermore, the pairwise TMD could be normalized by the maximum pairwise measure of 10.

Now consider the task of predicting if $Y \geq X$ proposed in the introduction. Figure 3 plots three decision trees, where the decision regions of Tree 1 and Tree 2 are the same as those in Figure 1, respectively. Supposing that we already have two decision trees, Tree 1 and Tree 3, and eight instances indicated in Figure 1, we want to add another decision tree from Tree 1 and Tree 2 to ensemble. None of the existing behavioral diversity measures can tell the difference between Tree 1 and Tree 2, since they predict the same on the eight instances. In contrast, tree matching diversity measure favors Tree 2 with a different structure rather than a duplicate of Tree 1. In fact, if X and Y are uniformly sampled from $[0, 10]$, the expected majority voting error by adding Tree 1 is 0.17 while that by adding Tree 2 is only 0.13. In practice, classifiers may predict quite differently, thus both structural and behavioral diversities should be considered together to get a good ensemble.

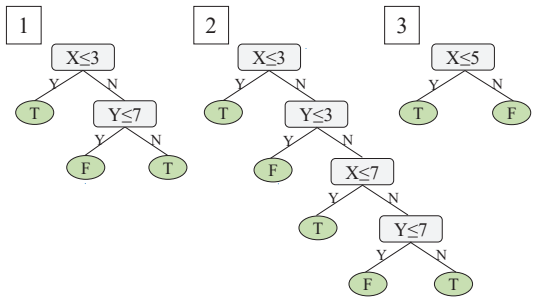


Fig. 3 Three decision trees for the task of predicting if $Y \geq X$. $X, Y \in [0, 10]$. Given Tree 1, Tree 3 and the eight instances in Figure 1, TMD (based on structural diversity) suggests to include Tree 2 rather than another Tree 1, whereas existing diversity measures (based on behavioral diversity) think there is no difference between Tree 1 and Tree 2.

4 Empirical Study

To investigate the usefulness of considering structural diversity for decision trees, we empirically evaluate performances of selective ensemble methods by incorporating different diversity measures with their enhancement by the tree matching diversity measure.

4.1 Selective Ensemble Methods

We take three selective ensemble methods as the platform for evaluating diversity measures. *Kappa Pruning* [27] is an ordering-based method, *Hierarchical Agglomerative Clustering (HAC) Pruning* [28] is a clustering-based method and *Semi-definite Programming (SDP) Pruning* [30] is an optimization-based method. For each method, different diversity measures are incorporated to replace the original ones. It should be noted that our intention in this paper is not to surpass the state-of-the-art selective ensemble methods, but to study the influence of incorporating structural diversity. Therefore the selected pruning methods are concise and all explicitly utilize the behavioral diversity measures.

The Kappa Pruning attempts to select the subset of most diverse classifiers. It measures diversity by the κ-statistic. The process initially selects the pair of classifiers with the lowest κ, then iteratively incorporates the classifier with the highest diversity with respect to the selected subensemble.

The HAC Pruning employs the hierarchical agglomerative clustering to partition the classifiers into several groups. The distance between two classifiers is defined as the probability they do not make coincident errors. Complete link method is used for the inter-cluster distance. Then for each cluster, one classifier with the maximum average distance from all other clusters is chosen as its prototype classifier. Finally, all the prototype classifiers are combined to form the pruned ensemble.

The SDP Pruning formulates ensemble pruning as a quadratic integer programming problem. Given T individual classifiers and N validation instances, the objective function can be equivalently stated as

$$\begin{aligned} &\min_{z} z^{\top}(\mathbf{G_{on}} + \mathbf{G_{off}})\, z \\ &s.t. \sum_{i=1}^{T} z_i = K,\ z_i \in \{0, 1\} \end{aligned} \tag{1}$$

where K is the predefined pruned ensemble size and z_i indicates whether classifier h_i is selected. $\mathbf{G_{on}}$ is a diagonal matrix, whose element $G_{on_{ii}}$ is the error made by h_i. $\mathbf{G_{off}}$ measures diversity, whose off-diagonal element $G_{off_{ij}}$ is the normalized common error made by h_i and h_j.

Different diversity measures are incorporated into these three approaches. For ordering-based methods, at each stage a classifier with the highest compared diversity with respect to the selected subensemble is included like Kappa Pruning. For clustering-based methods, the classifiers are first partitioned into several clusters as HAC Pruning does and then for each cluster one classifier with the maximum

Table 3 Comparison of test error and ensemble size (mean±std) when diversity measures and their TMD enhancement are used for ordering-based selective ensemble methods. Bold highlight the significantly smaller error/size of each pair of comparison, according to pairwise *t*-tests at significance level 0.95.

	κ	κ+TMD	DIS	DIS+TMD	ENT	ENT+TMD	TMD	BG
				Test Error				
anneal.ORIG	.113±.025	**.093±.021**	.107±.029	**.091±.021**	.095±.025	**.091±.022**	.090±.021	.094±.022
breast-cancer	.293±.029	.293±.027	.302±.034	**.290±.032**	.306±.032	**.292±.034**	.287±.034	.289±.032
breast-w	.049±.017	**.044±.014**	.049±.013	**.045±.013**	.044±.012	.044±.013	.044±.011	.048±.015
credit-a	.146±.020	**.142±.019**	.144±.018	.142±.019	.144±.019	.142±.019	.142±.019	.142±.018
kr-vs-kp	.020±.007	**.017±.006**	.020±.007	**.016±.006**	.018±.007	**.016±.006**	.017±.006	.017±.007
mfeat-factors	.079±.014	**.076±.013**	.080±.014	**.076±.013**	.075±.014	.076±.013	.076±.014	.081±.016
mfeat-fourier	.231±.017	.228±.016	.233±.017	**.228±.017**	.228±.016	.228±.018	.229±.017	.229±.015
page-blocks	.032±.004	.032±.004	.033±.004	.032±.004	.032±.004	.032±.004	.032±.004	.033±.004
phishing	.052±.003	.051±.004	.052±.004	**.051±.004**	.052±.004	**.051±.004**	.051±.004	.052±.003
phoneme	.152±.010	**.146±.009**	.152±.009	**.147±.009**	.151±.009	**.147±.009**	.147±.009	.150±.009
segment	.054±.010	**.050±.011**	.054±.010	**.050±.011**	.053±.010	**.050±.011**	.050±.011	.056±.011
seismic-bumps	.067±.002	.067±.002	.070±.005	**.069±.004**	.069±.004	.069±.004	.069±.004	.066±.001
sick	.019±.004	**.018±.004**	.018±.004	.018±.004	.018±.004	.018±.004	.018±.004	.019±.004
splice	.082±.011	**.078±.010**	.083±.011	**.079±.011**	.080±.011	.079±.010	.079±.011	.082±.011
texture	.061±.008	**.058±.008**	.061±.007	**.058±.007**	.059±.008	.058±.007	.058±.008	.060±.007
tic-tac-toe	.179±.019	**.174±.019**	.174±.026	.170±.021	.173±.025	.170±.022	.170±.021	.186±.028
twonorm	.047±.006	.046±.005	.047±.006	.046±.006	**.045±.005**	.046±.006	.046±.006	.042±.005
vehicle	.278±.027	.279±.026	.280±.028	.281±.027	.279±.028	.278±.027	.278±.026	.288±.022
vowel	.291±.037	**.277±.036**	.288±.037	**.278±.034**	.285±.038	**.278±.035**	.277±.035	.285±.037
wine-white	.427±.011	**.423±.010**	.427±.013	**.423±.012**	.425±.013	.423±.012	.423±.012	.419±.012
W/T/L	0/7/13		0/6/14		1/12/7		– –	– –
				Ensemble Size				
anneal.ORIG	16.4±7.0	16.4±7.1	15.8±6.6	15.6±5.1	16.5±5.3	16.1±5.6	14.9±4.7	100
breast-cancer	13.7±5.1	13.2±4.1	13.9±4.1	15.4±5.6	13.8±4.7	15.4±7.0	14.3±4.5	100
breast-w	14.1±5.0	13.5±4.1	13.5±4.6	14.1±4.5	14.4±4.1	14.0±3.9	14.7±5.0	100
credit-a	14.7±4.7	**12.9±3.5**	15.4±4.7	**13.4±4.3**	14.7±4.4	13.8±4.6	13.1±4.6	100
kr-vs-kp	16.3±5.5	**14.0±4.7**	14.7±5.8	14.7±5.3	16.3±6.8	14.4±5.1	14.9±5.7	100
mfeat-factors	21.9±5.7	22.8±6.7	20.6±5.7	22.4±6.5	22.6±4.9	22.3±6.4	23.1±6.7	100
mfeat-fourier	20.4±6.9	20.9±6.8	20.5±6.4	20.9±6.8	22.2±6.5	20.9±7.2	21.1±7.7	100
page-blocks	16.4±6.3	16.1±5.0	15.2±5.0	16.0±5.5	15.4±5.9	15.9±5.0	16.0±4.8	100
phishing	21.1±7.3	**18.9±6.5**	19.1±6.7	19.0±6.4	20.8±6.6	18.9±6.3	18.2±6.6	100
phoneme	25.0±8.2	**19.4±7.3**	22.6±9.1	**19.0±6.2**	21.9±7.6	**18.9±6.7**	18.7±7.0	100
segment	18.7±5.9	18.4±5.1	16.8±5.9	18.2±5.0	18.4±4.9	19.0±5.3	17.8±5.0	100
seismic-bumps	9.1±0.4	9.1±0.4	17.9±7.0	18.1±7.1	18.3±5.4	17.9±7.0	17.7±6.7	100
sick	13.7±4.7	13.4±4.5	14.4±4.5	**12.2±3.8**	13.3±3.7	**11.9±3.6**	12.5±4.4	100
splice	20.4±8.0	19.2±7.3	19.8±6.5	19.1±7.0	19.3±6.6	18.7±6.6	18.6±6.3	100
texture	26.3±9.1	25.8±6.5	**23.0±6.1**	26.0±6.5	26.3±7.7	25.7±6.5	25.9±6.7	100
tic-tac-toe	19.3±6.4	18.6±6.4	18.1±6.2	16.8±5.5	19.1±7.1	**16.6±4.7**	18.2±6.2	100
twonorm	37.7±8.6	39.0±11.1	36.8±9.3	38.8±11.8	**35.8±8.6**	39.2±12.5	37.7±11.3	100
vehicle	20.6±6.6	**19.0±4.3**	18.3±6.4	19.5±5.1	19.0±5.0	18.5±5.3	18.1±5.2	100
vowel	24.1±6.9	23.4±7.2	23.5±7.0	24.2±8.2	25.8±6.7	**23.4±7.1**	23.9±7.2	100
wine-white	**25.2±7.8**	28.8±8.6	25.6±8.3	27.8±9.1	27.9±9.2	27.5±8.5	27.7±8.5	100
W/T/L	1/14/5		1/16/3		1/15/4		– –	– –

compared diversity with respect to all other clusters is chosen as the prototype classifier. For optimization-based methods, the objective functions are similar to SDP Pruning except that the diversity matrix $\mathbf{G}_{\mathbf{off}}$ is substituted by the compared diversity matrix.

4.2 Configuration

We use 20 data sets for the experiments, in which 10 are binary-class and the other 10 are multi-class. The number of instances of the data sets ranges from 286 to 11055 and the dimension ranges from 5 to 216. Since they are commonly used data sets, details are left out. Each data set is randomly and evenly split for training, validating and testing. A Bagging of 100 J48 decision trees [35] is trained on the training set, then selected using the validation set, and finally tested on the test set. We repeat the above process for 10 times and for all 6 permutations of the split. Thus, the reported results are the average of 60 runs. For ordering-based and clustering-based methods, the size of the final selected ensemble is chosen by the validation set. For optimization-based methods, the size is set to 15, close to that chosen by the validation set on most data sets. To study the influence of incorporating structural diversity, each behavioral diversity measure in Table 2 is compared to its enhancement by the tree matching diversity measure. Before combination, all the diversity measures

Table 4 Comparison of test error and ensemble size (mean±std) when diversity measures and their TMD enhancement are used for clustering-based selective ensemble methods. Bold highlight the significantly smaller error/size of each pair of comparison, according to pairwise *t*-tests at significance level 0.95.

	κ	κ+TMD	DIS	DIS+TMD	ENT	ENT+TMD	TMD	BG
Test Error								
anneal.ORIG	.075±.018	**.070±.015**	.073±.017	**.070±.015**	.073±.016	**.070±.015**	.070±.015	.094±.022
breast-cancer	.287±.029	.286±.027	.290±.031	**.283±.029**	.290±.031	**.283±.029**	.284±.032	.289±.032
breast-w	.041±.014	.040±.014	.040±.013	.040±.014	.040±.013	.040±.014	.040±.014	.048±.015
credit-a	.143±.018	**.140±.020**	.143±.018	**.140±.019**	.143±.018	**.140±.019**	.139±.019	.142±.018
kr-vs-kp	.011±.004	.011±.003	.011±.004	.011±.003	.011±.004	.011±.003	.011±.003	.017±.007
mfeat-factors	.073±.013	.074±.012	.073±.012	.074±.012	.073±.013	.074±.012	.073±.012	.081±.016
mfeat-fourier	.224±.017	.224±.015	.224±.015	.224±.014	.224±.016	.224±.015	.224±.014	.229±.015
page-blocks	.031±.004	.031±.004	.031±.003	.031±.004	.031±.004	.031±.004	.031±.004	.033±.004
phishing	.051±.003	.050±.004	.050±.004	.050±.004	.050±.004	.050±.004	.050±.004	.052±.003
phoneme	.144±.008	.144±.008	.143±.009	.144±.009	.143±.009	.144±.009	.144±.009	.150±.009
segment	.049±.010	.049±.010	.049±.010	.049±.010	.049±.010	.049±.010	.049±.009	.056±.011
seismic-bumps	.067±.002	.067±.002	.068±.004	**.067±.002**	.068±.004	**.067±.002**	.067±.002	.066±.001
sick	.017±.004	.016±.003	.017±.003	**.016±.003**	.017±.003	**.016±.003**	.017±.004	.019±.004
splice	.075±.009	.075±.010	.076±.010	**.075±.010**	.075±.009	.075±.010	.075±.009	.082±.011
texture	.058±.007	.057±.007	.057±.007	.057±.007	.058±.007	.057±.007	.057±.007	.060±.007
tic-tac-toe	.164±.018	**.160±.018**	.163±.023	.160±.022	.163±.023	.160±.022	.160±.022	.186±.028
twonorm	.042±.004	.041±.004	.042±.004	.042±.004	.042±.004	.042±.004	.042±.004	.042±.005
vehicle	.274±.026	.274±.024	.277±.024	.274±.024	.276±.025	.274±.025	.274±.025	.288±.022
vowel	.268±.035	.268±.031	.268±.034	.268±.031	.269±.035	.268±.031	.269±.031	.285±.037
wine-white	.417±.011	.418±.012	.419±.012	.418±.012	.418±.012	.417±.012	.417±.012	.419±.012
W/T/L	0/17/3		0/14/6		0/15/5		– –	– –
Ensemble Size								
anneal.ORIG	23.8±11.8	**18.8±10.4**	20.2±12.4	19.6±12.1	20.1±12.2	19.6±12.1	18.5±9.9	100
breast-cancer	24.6±17.4	23.8±17.8	23.9±18.1	19.9±17.2	23.9±18.1	19.9±17.2	18.6±15.8	100
breast-w	15.6±10.4	14.1±8.5	14.9±10.2	14.2±8.6	14.9±10.2	14.2±8.6	14.1±8.9	100
credit-a	21.1±18.5	**17.8±15.7**	20.9±19.5	19.2±18.7	20.9±19.5	19.2±18.7	19.4±18.6	100
kr-vs-kp	20.0±16.3	**16.1±13.6**	20.1±16.3	**16.1±13.6**	20.1±16.3	**16.1±13.6**	16.3±13.6	100
mfeat-factors	28.6±14.9	28.4±14.0	26.3±12.6	28.0±14.5	29.6±15.1	27.1±14.3	28.8±15.7	100
mfeat-fourier	31.4±14.8	28.3±15.8	**24.6±12.7**	28.1±15.6	32.1±15.2	**28.1±15.6**	29.8±17.1	100
page-blocks	22.9±17.9	19.1±15.9	19.2±16.0	18.9±14.3	20.6±16.4	17.6±11.6	17.9±11.5	100
phishing	29.7±18.5	**24.4±14.4**	30.6±18.2	**24.4±14.4**	30.6±18.2	**24.4±14.4**	24.7±14.4	100
phoneme	33.1±14.8	**27.5±16.6**	33.4±14.8	**25.4±14.8**	33.4±14.8	**25.4±14.8**	26.6±15.3	100
segment	20.0±9.9	**15.6±9.9**	19.1±11.0	**15.6±9.9**	19.9±10.0	**15.6±9.9**	15.2±9.4	100
seismic-bumps	13.6±18.7	13.2±17.4	21.0±16.7	19.7±16.4	21.0±16.7	19.7±16.4	19.8±16.6	100
sick	15.7±12.5	16.8±13.7	17.9±12.3	17.2±13.8	17.9±12.3	17.2±13.8	17.4±14.0	100
splice	22.1±9.8	21.3±11.9	23.7±15.7	21.7±11.3	22.6±11.0	21.5±11.3	20.1±9.6	100
texture	32.3±11.2	31.8±12.0	**29.0±10.3**	32.2±12.2	32.9±11.1	32.1±11.9	32.5±11.9	100
tic-tac-toe	28.2±15.5	**23.6±14.2**	24.2±14.3	**20.8±13.6**	24.2±14.3	**20.8±13.6**	21.4±13.0	100
twonorm	61.0±15.5	64.4±16.1	61.4±15.4	64.1±17.3	61.4±15.4	64.1±17.3	63.5±17.7	100
vehicle	24.6±12.2	24.7±13.2	22.8±11.9	23.4±14.2	24.1±12.1	23.9±14.7	23.8±14.9	100
vowel	29.7±13.6	28.6±13.8	27.1±12.8	28.0±12.5	29.0±12.7	27.4±12.4	27.2±11.1	100
wine-white	42.1±17.1	43.7±16.3	39.3±17.2	42.6±16.1	38.8±15.0	41.6±16.7	43.3±17.4	100
W/T/L	0/13/7		2/13/5		0/14/6		– –	– –

are first normalized to have a maximum possible value of 1. Then those measures which increase as diversity grows are replaced with 1 minus the measures. For ordering-based and clustering-based methods, the harmonic mean of the behavioral diversity measure and the tree matching diversity measure is used, as it is more sensitive to small values, thus avoiding getting both medium behavioral and structural diversities. For optimization-based methods, the arithmetic mean is applied, as it is natural to add the two diversity matrices in the objective functions.

4.3 Results

In our experiments, we compare the original diversity measures to their TMD enhancement under three types of selective ensemble approaches. The comparison results are exhibited in Tables 3-5 respectively. In general, applying TMD enhancement, which takes the structural diversity into account, leads to a better or comparable performance. The results of the unpruned decision forests of Bagging (BG) are also listed in the tables.

Table 3 shows the comparison results for ordering-based selective ensemble methods. As can be seen, for all three behavioral diversity measures, their enhancement by TMD achieve significantly better performance on most data sets.

Table 5 Comparison of test error (mean±std) when diversity measures and their TMD enhancement are used for optimization-based selective ensemble methods (ensemble size fixed). Bold highlight the significantly smaller error of each pair of comparison, according to pairwise *t*-tests at significance level 0.95.

	κ	κ+TMD	DIS	DIS+TMD	ENT	ENT+TMD	TMD	BG
anneal.ORIG	.114±.022	**.096±.022**	.093±.023	.090±.022	.094±.022	**.090±.022**	.088±.020	.094±.022
breast-cancer	.295±.012	.293±.019	.299±.034	**.289±.033**	.299±.034	**.289±.033**	.288±.033	.289±.032
breast-w	.046±.013	**.045±.012**	.045±.013	.044±.012	.045±.013	.044±.012	.044±.013	.048±.015
credit-a	.144±.017	.143±.018	.144±.017	**.141±.018**	.144±.017	**.141±.018**	.141±.017	.142±.018
kr-vs-kp	.018±.006	**.017±.007**	.018±.006	**.016±.006**	.018±.006	**.016±.006**	.016±.006	.017±.007
mfeat-factors	.078±.013	.078±.012	**.075±.013**	.077±.013	.078±.013	.079±.012	.079±.013	.081±.016
mfeat-fourier	.231±.014	.230±.016	.228±.015	.228±.015	.231±.014	.231±.015	.230±.016	.229±.015
page-blocks	.032±.003	**.032±.003**	.032±.003	.032±.004	.032±.003	.032±.004	.032±.004	.033±.004
phishing	.052±.004	.051±.003	.052±.004	.052±.003	.052±.004	.052±.003	.052±.003	.052±.003
phoneme	.156±.014	**.147±.009**	.151±.010	**.146±.008**	.151±.010	**.146±.008**	.146±.009	.150±.009
segment	.053±.010	**.049±.009**	.052±.010	**.049±.010**	.053±.010	**.050±.010**	.050±.010	.056±.011
seismic-bumps	**.066±.001**	.066±.001	.070±.004	.069±.004	.070±.004	.069±.004	.069±.004	.066±.001
sick	.017±.004	.018±.004	.018±.004	.018±.004	.018±.004	.018±.004	.018±.004	.019±.004
splice	.080±.010	**.077±.009**	.080±.010	**.078±.009**	.079±.010	**.078±.009**	.079±.010	.082±.011
texture	.061±.008	**.060±.008**	.061±.008	**.060±.008**	.061±.008	**.060±.007**	.060±.008	.060±.007
tic-tac-toe	.184±.021	**.172±.018**	.166±.022	.164±.021	.166±.022	.164±.021	.171±.024	.186±.028
twonorm	.057±.005	**.056±.005**	.057±.005	.057±.005	.057±.005	.057±.005	.057±.005	.042±.005
vehicle	.276±.028	.278±.024	.280±.025	.275±.027	.277±.027	.276±.024	.281±.028	.288±.022
vowel	.299±.036	**.282±.035**	.281±.034	.277±.032	.298±.034	**.283±.035**	.287±.033	.285±.037
wine-white	.433±.012	**.431±.013**	.431±.012	.430±.013	.433±.012	.431±.014	.431±.012	.419±.012
W/T/L	1/7/12		1/12/7		0/11/9		– –	– –

Specifically, from the t-test with significance level 0.95, κ+TMD wins κ on 13 data sets in test errors. DIS+TMD wins DIS on 14 data sets in test errors. Although ENT+TMD wins ENT on only 7 data sets in test errors, the ensemble sizes of ENT+TMD are much smaller than those of ENT. Table 4 shows the comparison results for clustering-based methods. In this case, the improvement in test error by combining TMD is not as impressive as that in ordering-based methods. The reason might be that the same partitions of the base learners are used for all diversity measures and the difference lies only in choosing cluster prototype classifiers. However, the results are still encouraging as the enhancement by TMD achieves lower test errors with much smaller ensemble sizes. The results for optimization-based methods are listed in Table 5. Similar results can be observed that the TMD enhancement achieves a better or comparable performance. TMD alone performs pretty well. Since the local structure of decision trees characterizes the partitions of the feature space, TMD would be more efficient in measuring the difference than those behavioral diversity measures based on finite instances. Still, the TMD enhancement performs slightly better than TMD alone, indicating the necessity to consider both diversity measures together.

The overall W/T/L counts of the compared diversity measures across three types of selective ensemble methods are summarized in Table 6. For all three diversity measures, the enhancement by TMD reduces test errors and ensemble sizes significantly. These empirical results validate that it is useful to consider both behavioral and structural diversity. Here we do not make parameter selection for explicitness and the combination techniques are primitive. Better performance of the combination is expected by delicate designs.

Table 6 Summary of W/T/L counts of TMD enhancement across three types of selective ensemble methods.

	κ vs. κ+TMD	DIS vs. DIS+TMD	ENT vs. ENT+TMD
Test error	1/31/28	1/32/27	1/38/21
Ensemble size	1/27/12	3/29/8	1/29/10

To better compare the performance of the selective ensemble methods with different diversity measures, we perform *Friedman test* in conjunction with the *Bonferroni-Dunn test* at significance level 0.95, which is a non-parametric statistical significance test for comparing multiple classifiers on multiple data sets based on the ranks. Figure 4 shows the results of Friedman test of the compared diversity measures in three catogeries of selective ensemble methods. As can be seen, for all three behavioral diversity measures, their TMD enhancement perform overally better on 20 data sets. The conclusions agree with the t-test results.

4.4 Training time

Table 7 gives the accumulated training time of the compared selective ensemble methods on 20 data sets. The experiments are performed on a PC with 2.0 GHz CPU and 8 GB memory, implemented by MATLAB. In the table, the x column is

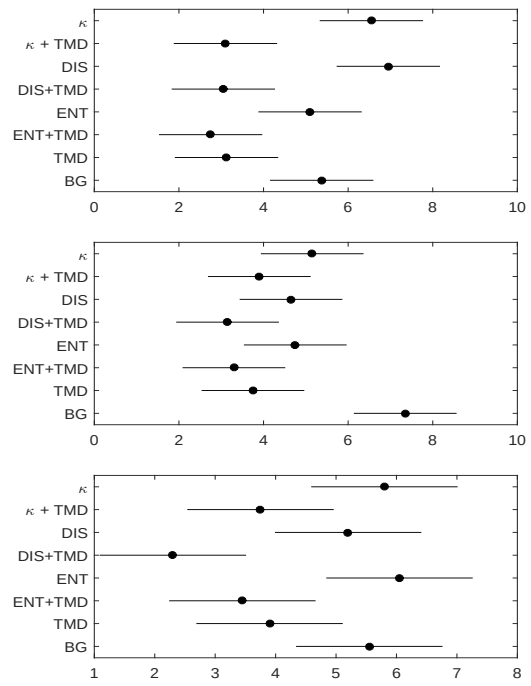


Fig. 4 Results of the Friedman test for comparing the performance of selective ensemble with different diversity measures on 20 data sets. The dots indicate the average ranks; the bars indicate the critical difference with the Bonferroni-Dunn test at significance level 0.95. From top to bottom: ordering-based, clustering-based and optimization-based selective ensemble methods.

averaged over κ, DIS and ENT, and the x+TMD column is averaged over κ+TMD, DIS+TMD and ENT+TMD. The extra training time of x+TMD compared to x is mainly consumed on the computation of pairwise structural diversity matrices.

Table 7 Accumulated training time (seconds, mean ± std) on 20 data sets for three categories of selective ensemble methods, repeated by 10 times. x and x+TMD are averaged over three diversity measures, κ, DIS and ENT.

	x	x+TMD	TMD
ordering	496.78±0.37	593.12±1.40	92.60±1.43
clustering	88.97±0.26	219.89±1.48	161.71±1.65
optimization	23.53±0.34	115.47±1.67	99.42±1.70

4.5 Influence of the Initial Pool Size

In the previous experiments, the initial pool size, i.e., the number of base learners in the full ensemble, is set to 100. To investigate the influence of the initial pool size, a Bagging of 500, 1000 and 1500 J48 selective decision tree ensembles are evaluated, respectively.

Figure 5 shows that no matter what size of the initial pool, the TMD enhanced versions are always better than their corresponding original versions. Here we only plot the results on *phoneme* (binary classification) and *aneal.ORIG* (multi-class classification), whereas most other data sets are with similar tendencies.

4.6 Decision Forests with Random Subspace

Bagging uses boostrap samples to build different trees. To further explore the usefulness of considering behavioral and structural diversities together, the Random Subspace is used to train the 100 J48 decision trees. By the default settings of Weka [35] , only a random half of features are selected to train each decision tree. Thus, the decision trees built with Random Subspace are expected to be more diverse in their structure than those built with Bagging. Followed by the same experiment procedures as before, the three behavioral diversity measures and their TMD enhancement are used to prune the 100 decision trees. Table 8 summarizes the overall W/T/L counts of the compared diversity measures across three types of selective ensemble methods with Random Subspace. It can be seen that the enhancement by TMD reduces test errors and ensemble sizes significantly, in accord with the results with Bagging.

Table 8 Summary of W/T/L counts of TMD enhancement across three types of selective ensemble methods with Random Subspace.

	κ vs. κ+TMD	DIS vs. DIS+TMD	ENT vs. ENT+TMD
Test error	5/24/31	7/34/19	6/31/23
Ensemble size	3/26/11	4/32/4	3/27/10

5 Conclusion

In this paper, we advocate to consider structural diversity in addition to behavioral diversity for ensemble methods. We proposed the tree matching diversity measure for decision tree. The experiments on selective ensemble validate the necessity to consider both structural and behavioral diversities together. The proposed measure could be used in other decision tree ensemble learning methods.

This paper considers only decision trees, and it will be interesting to design structural diversity for other kinds of base learners. Even for decision trees, this paper considers a simple definition of structural diversity, and it is interesting to explore better alternatives. Also, how to exploit behavioral and structural diversities together is also worth studying in

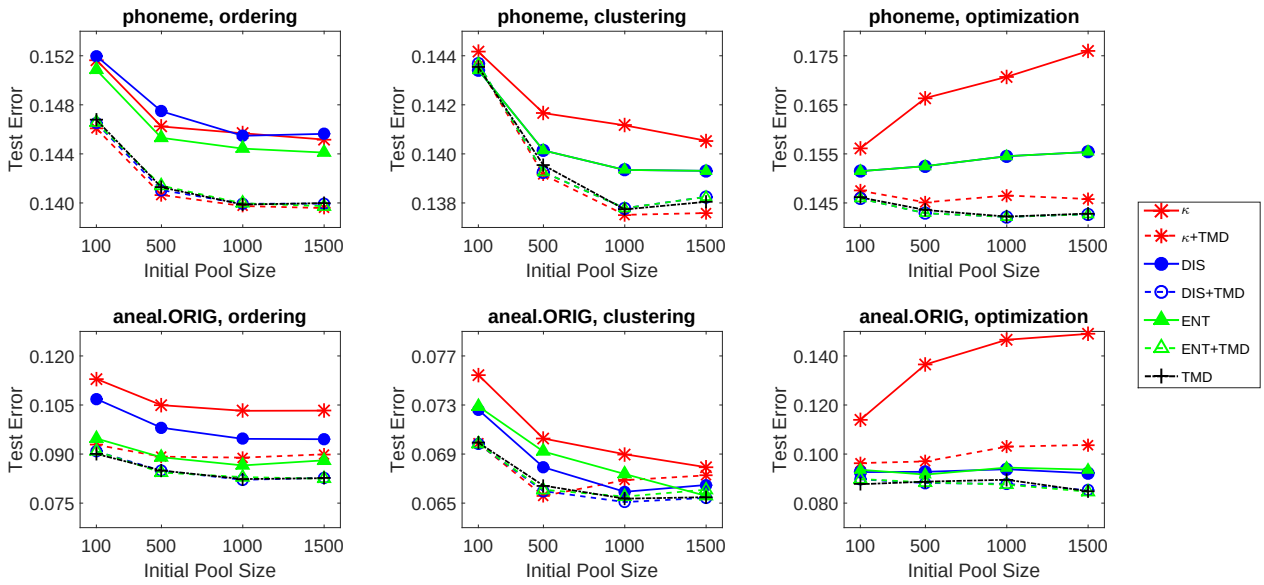


Fig. 5 Influence of initial pool size on test error. Increasing pool size does not necessarily lead to better performance. However, TMD enhanced versions (dash lines) are superior to original versions (solid lines).

the future.

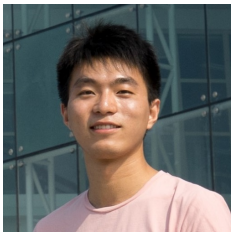
Acknowledgements The authors would like to thank anonymous reviewers for their helpful comments and suggestions. This research was supported by the NSFC (61333014).

References

1. Stiglic G, Kocbek S, Pernek I, and Kokol P. Comprehensive decision tree models in bioinformatics. *PloS One*, 7(3):e33812, 2012.
2. Creamer G and Freund Y. Using boosting for financial analysis and performance prediction: application to s&p 500 companies, latin american adrs and banks. *Computational Economics*, 36(2):133–151, 2010.
3. Rokach L. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 2016.
4. Zhou Z H. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, Boca Raton, FL, 2012.
5. Breiman L. Random forests. *Machine Learning*, 45(1):5–32, 2001.
6. Geurts P, Ernst D, and Wehenkel L. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.
7. Rodriguez J J, Kuncheva L I, and Alonso C J. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.
8. Brown G, Wyatt J, Harris R, and Yao X. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.
9. Melville P and Mooney R J. Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 505–510, 2003.
10. Yu Y, Li Y F, and Zhou Z H. Diversity regularized machine. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1603–1608, 2011.
11. Breiman L. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40(3):229–242, 2000.
12. Kuncheva L I and Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.
13. Tang E K, Suganthan P N, and Yao X. An analysis of diversity measures. *Machine Learning*, 65(1):247–271, 2006.
14. Didaci L, Fumera G, and Roli F. Diversity in classifier ensembles: Fertile concept or dead end? In *Proceedings of the 11th International Workshop on Multiple Classifier Systems*, pages 37–48, 2013.
15. Reyzin L and Schapire R E. How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 753–760, 2006.
16. Quinlan J R. Simplifying decision trees. *International Journal of Human-Computer Studies*, 51(2):497–510, 1999.
17. Freund Y and Mason L. The alternating decision tree learning algorithm. In *Proceedings of the 16th International Conference on Machine Learning*, volume 99, pages 124–133, 1999.
18. Friedman J H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.
19. Krogh A and Vedelsby J. Neural network ensembles, cross validation, and active learning. In *Advances In Neural Information Processing Systems 7*, pages 231–238, 1995.
20. Geman S, Bienenstock E, and Doursat R. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.
21. Margineantu D D and Dietterich T G. Pruning adaptive boosting. In *Proceedings of the 14th International Conference on Machine*

Learning, pages 211–218, 1997.

22. Brown G and Kuncheva L I. “Good" and “bad" diversity in majority vote ensembles. In *Proceedings of the 9th International Workshop on Multiple Classifier Systems*, pages 124–133, 2010.
23. Brown G. An information theoretic perspective on multiple classifier systems. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 344–353, 2009.
24. Zhou Z H and Li N. Multi-information ensemble diversity. In *Proceedings of the 9th International Workshop on Multiple Classifier Systems*, pages 134–144, 2010.
25. Brooks F P Jr. Three great challenges for half-century-old computer science. *Journal of the ACM*, 50(1):25–26, 2003.
26. Zhou Z H, Wu J, and Tang W. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1):239–263, 2002.
27. Martínez-Muñoz G, Hernández-Lobato D, and Suárez A. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):245–259, 2009.
28. Giacinto G, Roli F, and Fumera G. Design of effective multiple classifier systems by clustering of classifiers. In *Proceedings of the 15th International Conference on Pattern Recognition*, pages 160–163, 2000.
29. Lazarevic A and Obradovic Z. Effective pruning of neural network classifier ensembles. In *2001 International Joint Conference on Neural Networks*, pages 796–801, 2001.
30. Zhang Y, Burer S, and Street W N. Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research*, 7:1315–1338, 2006.
31. Li N and Zhou Z H. Selective ensemble under regularization framework. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 293–303, 2009.
32. Qian C, Yu Y, and Zhou Z H. Pareto ensemble pruning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2935–2941, 2015.
33. Pawlik M and Augsten N. Tree edit distance: Robust and memory-efficient. *Information Systems*, 56:157–173, 2016.
34. Wolberg W H and Mangasarian O L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23):9193–9196, 1990.
35. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten I H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.



Tao Sun received his BS degree in the School of Automation from Huazhong University of Science and Technology, China, in 2015. He is currently a graduate student in the Department of Computer Science and Technology, Nanjing University, China. His research interests include machine learning and data mining.

Zhi-Hua Zhou is a professor at the Department of Computer Science and Technology, Nanjing University, China. He is the founding director of LAMDA. He is a foreign member of the Academy of Europe, and fellow of the ACM, AAAI, AAAS, IEEE, IAPR, and CCF. His main research interests are in artificial intelligence, machine learning and data mining.