Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 3845131, 11 pages
http://dx.doi.org/10.1155/2016/3845131

Research Article

A Novel Margin-Based Measure for Directed Hill Climbing Ensemble Pruning

Huaping Guo,[1] Fang Sun,[1] Jiong Cheng,[1] Yanling Li,[1] and Mingling Xu[2]

[1]*School of Computer and Information Technology, Xinyang Normal University, Xinyang, Henan 464000, China*
[2]*School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450000, China*

Correspondence should be addressed to Huaping Guo; hpguo_cm@163.com

Received 10 February 2016; Revised 14 June 2016; Accepted 30 June 2016

Academic Editor: Simone Bianco

Ensemble pruning is a technique to increase ensemble accuracy and reduce its size by choosing a subset of ensemble members to form a subensemble for prediction. Many ensemble pruning algorithms via directed hill climbing searching policy have been recently proposed. The key to the success of these algorithms is to construct an effective measure to supervise the search process. In this paper, we study the importance of individual classifiers with respect to an ensemble using margin theory proposed by Schapire et al. and obtain that ensemble pruning via directed hill climbing strategy should focus more on examples with small absolute margins as well as classifiers that correctly classify more examples. Based on this principle, we propose a novel measure called the margin-based measure to explicitly evaluate the importance of individual classifiers. Our experiments show that using the proposed measure to prune an ensemble leads to significantly better accuracy results compared to other state-of-the-art measures.

1. Introduction

Ensemble of multiple learning machines has been a very popular research topic during the last decade in machine learning and data mining. The basic idea is to construct multiple classifiers from the original data and then aggregate their predictions when classifying examples with unknown classes. Theoretic and empirical results show that an ensemble is potential to increase the classification accuracy beyond the level reached by an individual classifier alone [1]. Dietterich stated "A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse" [2].

Many approaches have been proposed to create ensemble members with both high accuracy and high diversity, which can be mainly grouped into three categories: (1) by manipulating data set [3, 4], (2) by manipulating features [5–8], and (3) by manipulating algorithms [9]. Bagging [3] and boosting [4], the most widely used and successful ensemble learning methods, fall into the first category, where bagging learns individual classifiers on data sets obtained by randomly sampling from the original training sets and, through randomly disturbing, the learned classifiers obtain a high accuracy and sufficient diversity. Unlike bagging, boosting is an iterative learning process. For each iteration, boosting adjusts the distribution of training set such that classifiers focus more on examples that are hardly correctly classified. The approaches by manipulating features try to build the individual classifiers on diverse feature spaces obtained by selecting subset or by generating new ones from the original features. For example, random forests [5, 6] learn each tree on a feature subset obtained by randomly sampling from original features and COPEN [8] learns the base classifiers on new feature spaces mapped from original feature space using pairwise constraints projection. The individual classifiers can also be built by manipulating algorithms. Through adjusting model structure or parameter setting, classifiers with diversity are learned, such that the negative correlation method explicitly constrains the parameters of individual neural networks to be different by a regularization term [9].

Ensemble methods have been successfully applied to many fields such as remote sensing [10], time series prediction [11], and imbalanced learning problem [12]. However, an obvious problem existing in ensemble learning methods is that they tend to train a very large number of classifiers which need large storage resources to store them and computational

resources to calculate outputs of individual learners. Besides, it is not always true that the larger the ensemble, the better its performance. In fact, Zhou et al. [13] proved that the generalization performance of a subset of an ensemble may be even better than the ensemble consisting of all the given individual learners. These reasons motivate the appearance of ensemble pruning, also called ensemble selection or ensemble thinning, selecting a subset of ensemble members to form subensembles that are subject to less resource consumption and response time with accuracy that is similar to or better than the original ensemble [14–22].

Given an ensemble with M members, searching for the best subset of ensemble members by enumerating all subensemble candidates is computational infeasible because of exponential size of the search space $2^M - 1$, which is NP-complete problem [23]. Several efficient methods that are based on a directed hill climbing search in the space of subsets report good predictive performance results [15, 16, 18, 24–27]. These methods start with an empty (or full) initial ensemble and search the space of different ensembles by iteratively expanding (or contracting) the initial ensemble by a single model. The search is guided by an evaluation measure that is based on either the predictive performance or the diversity of the alternative subsets. The evaluation measure is the main component of a directed hill climbing algorithm and it differentiates the methods that fall into this category.

In this paper, we apply the concepts of example margins proposed by Schapire et al. [28] to analyse the importance of individual classifiers with respect to an ensemble and conclude that ensemble pruning via directed hill climbing strategy should focus more on examples with small absolute margins as well as classifiers that correctly classify more examples. Based on the gained insight, a criterion called margin-based measure is proposed to supervise the search process of ensemble pruning via directed hill climbing strategy. Our experiments show that using the proposed measure to prune an ensemble leads to significantly better accuracy results compared to other state-of-the-art measures.

The paper is structured as follows. Section 2 briefly describes ensemble pruning via directed hill climbing search. Section 3 proposes a measure for evaluating the importance of individual classifiers. Section 4 reports the experimental settings and results, and we conclude this paper in Section 5.

2. Related Work

Directed hill climbing ensemble pruning (DHCEP) attempts to find the globally best subset of classifiers by taking local greedy decisions for changing the current subset [17, 28, 29]. An example of the search space for an ensemble of four models is presented in Figure 1.

The direction of search and the measure used for evaluating the search are two important parameters that differentiate one DHCEP method from the other. The following sections discuss the different options for instantiating these parameters and the particular choices of existing methods.

2.1. Direction of Search. Based on the direction of search we have two main categories of DHCEP methods: (a) forward selection and (b) backward elimination (see Figure 1).

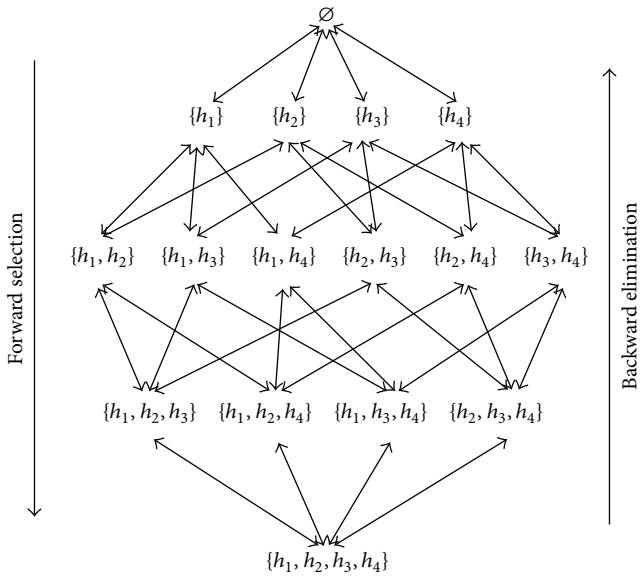

Figure 1: The search space of DHCEP methods for an ensemble of 4 models.

In forward selection algorithm, ensemble pruning starts with the current classifier subset S which is initialized to the empty set. Then the algorithm continues by iteratively adding to S the classifier $h \in H \setminus S$ that optimizes an evaluation function. This function evaluates the addition of classifier h in the current subset S based on the pruning set (labeled data). In the past, this approach has been used in [14, 25, 26] and in reduce-error pruning methods [30, 31].

In backward elimination, the current classifier subset S is initialized to the complete ensemble H and the algorithm continues by iteratively removing from S the classifier $h \in S$ that optimizes the evaluation function. This function evaluates the removal of classifier h from the current subset S based on the pruning set. In the past, this approach has been used in the AID thinning and concurrency thinning algorithms [15].

In both cases, the traversal requires the evaluation of $M(M + 1)/2$ subsets, leading to a time complexity of $O(M^2 g(M, N))$, where the term $g(M, N)$ concerns the complexity of the evaluation function, which is linear with respect to N (the size of pruning set) and ranges from constant to quadratic with respect to M (the size of H), as we will see in the following sections.

2.2. Evaluation Measure. Evaluation measures are the main component that differentiates DHCEP methods, which can be grouped into two major categories: those are based on performance and those are based on diversity.

The goal of performance-based measures is to find the model that maximizes the performance of the ensemble produced by adding (or removing) a model to (or from) the current ensemble. Their calculation depends on the method used for ensemble combination, which usually is voting. Accuracy was used as an evaluation measure by Margineantu and Dietterich [30] and by Fan et al. [25], while Caruana et al. [26] experimented with several measures, including

accuracy, root mean squared error, mean cross-entropy, lift, precision/recall break-even point, precision/recall F-score, average precision, and ROC area. Another measure is benefit, which is based on a cost model and has been used in Fan et al. [25]. The calculation of performance-based metrics requires the decision of the ensemble on all examples of the pruning set. Therefore, the complexity of these measures is $O(|S|N)$. However, this complexity can be optimized to $O(N)$, if the predictions of the current ensemble are updated incrementally each time a classifier is added to (or removed from) it.

Ensemble diversity, that is, the difference among the individual learners, is a fundamental issue in ensemble methods. Intuitively, it is easy to understand that, in order to gain from a combination, individual learners must be different, and otherwise there would be no performance improvement if identical individual learners were combined.

Let h be a classifier and let S be subensemble; Partalas et al. [16, 18, 29] identify that the prediction of h and S on an instance $\mathbf{x}_i$ can be categorized into four cases: (1) e_{tt}: $h(\mathbf{x}_i) = y_i \wedge S(\mathbf{x}_i) = y_i$, (2) e_{ft}: $h(\mathbf{x}_i) \neq y_i \wedge S(\mathbf{x}_i) = y_i$, (3) e_{tf}: $h(\mathbf{x}_i) = y_i \wedge S(\mathbf{x}_i) \neq y_i$, and (4) e_{ff}: $h(\mathbf{x}_i) \neq y_i \wedge S(\mathbf{x}_i) \neq y_i$. They concluded that considering the four cases is crucial to design ensemble diversity measure. Many diverse measures are designed by considering some or all the four cases, for example, complementariness [14] and concurrency [15]. The complementariness of h with respect to S and a pruning set D_{pr} is calculated as

$$\text{COM}\left(h, S, D_{\text{pr}}\right) = \sum_{\mathbf{x}_i \in D_{\text{pr}}} I\left(\mathbf{x}_i \in e_{tf}\right), \tag{1}$$

where I (true) $= 1$, I (false) $= 0$. The complementariness is exactly the number of examples that are correctly classified by h and incorrectly classified by S. The concurrency is defined as

$$\begin{aligned}&\text{CON}\left(h, s, D_{\text{pr}}\right)\\ &\quad= \sum_{\mathbf{x}_i \in D_{\text{pr}}} \left\{2I\left(\mathbf{x}_i \in e_{tf}\right) + I\left(\mathbf{x}_i \in e_{tt}\right) - 2I\left(\mathbf{x}_i \in e_{ff}\right)\right\}\end{aligned} \tag{2}$$

which is similar to the complementariness, with the difference that it considers two more cases and weights them.

Unlike complementariness and concurrency, Partalas et al. [18] introduce a new metric called uncertainty weighted accuracy (UWA) considering all four cases given above. UWA is defined as

$$\begin{aligned}\text{UWA}\left(h, S, D_{\text{pr}}\right) = \sum_{\mathbf{x}_i \in D_{\text{pr}}} &\left\{\text{NT}_i I\left(\mathbf{x}_i \in e_{tf}\right)\right.\\ &+ \text{NF}_i I\left(\mathbf{x}_i \in e_{tt}\right) - \text{NF}_i I\left(\mathbf{x}_i \in e_{ft}\right)\\ &\left.- \text{NT}_i I\left(\mathbf{x}_i \in e_{ff}\right)\right\},\end{aligned} \tag{3}$$

where NT_i is the proportion of classifiers in the current ensemble S which correctly predict $\mathbf{x}_i$ and $\text{NF}_i = 1 - \text{NT}_i$ is the proportion of classifiers that incorrectly predict $\mathbf{x}_i$. In addition to considering all four cases, UWA takes into account the strength of the decision of the current ensemble.

In this paper, we designed a new measure by considering the margin of examples for ensemble pruning via directed hill climbing. More details are discussed in next section.

3. Importance Assessment for Individual Classifiers

As one of the best off-the-shelf algorithms, AdaBoost demonstrates a high generalization performance. To theoretically analyse this phenomenon, a concept called margin of examples was proposed by Schapire et al. [28]. Let $D = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \ldots, N\}$ be the training set, where each example $\mathbf{x}_i$ is associated with a label $y_i \in \{-1, +1\}$. Suppose that $H = \{h_i \mid i = 1, 2, \ldots, M\}$ is an ensemble with M classifiers and suppose that each member $h \in H$ maps each example $\mathbf{x}_i \in D$ to a label y; namely, $h : \mathbf{x}_i \to y \in \{-1, 1\}$. Then the margin of $\mathbf{x}_i$ is defined as

$$\text{margin}\left(\mathbf{x}_i\right) = \frac{y_i \sum_{j=1}^{M} w_j h_j\left(\mathbf{x}_i\right)}{\sum_{j=1}^{M} w_j}, \tag{4}$$

where w_j is the weight of the classifier h_j. Without loss of generality, normalizing w_j, $j = 1, 2, \ldots, M$, such that $\sum w_j = 1$, then (4) can be written as

$$\text{margin}\left(\mathbf{x}_i\right) = y_i \sum_{j=1}^{M} w_j h_j\left(\mathbf{x}_i\right). \tag{5}$$

From (5), the margin is a value in $[-1, 1]$, $\mathbf{x}_i$ is on the border if $\text{marin}(\mathbf{x}_i) = 0$, the absolute value of the margin is the confidence of ensemble prediction on $\mathbf{x}_i$, and $\text{margin}(\mathbf{x}_i) > 0$ (or $\text{margin}(\mathbf{x}_i) < 0$) indicates that the ensemble correctly (or incorrectly) classifies $\mathbf{x}_i$. Based on this concept, they proved that, for any $\theta > 0$ and $\delta > 0$, the generalization error is upper bound by

$$\begin{aligned}&\widehat{\Pr}\left[\text{margin}\left(\mathbf{x}\right) \leq \theta\right]\\ &\quad+ \widehat{O}\left(\frac{1}{\sqrt{N}}\left(\frac{\ln N \ln d}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2}\right),\end{aligned} \tag{6}$$

where d is the complex of the base classifier and N is the size of the training set. To further explain the correctness of the margin theory, Gao and Zhou [32] proposed kth margin theory. Specifically, for any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of training set with size $N \geq 5$, the generalization error is upper bound by

$$\begin{aligned}&\frac{2}{N} + \inf_{\theta \in (0,1]} \left\{\widehat{\Pr}\left[\text{margin}\left(\mathbf{x}\right) \leq \theta\right] + \frac{7\mu + 3\sqrt{3\mu}}{3N}\right.\\ &\quad\left.+ \sqrt{\frac{3\mu}{N}\widehat{\Pr}\left[\text{margin}\left(\mathbf{x}\right) \leq \theta\right]}\right\},\end{aligned} \tag{7}$$

where

$$\mu = \frac{8}{\theta^2}\ln N \ln\left(2d\right) + \ln\frac{2d}{\delta}. \tag{8}$$

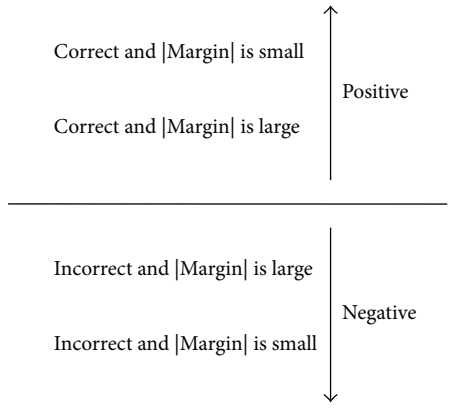

Figure 2: Rules obtained from the margin theory for evaluating the importance of individual classifiers on examples.

From (6) and (7), when other variables are fixed, the larger the margin over the training examples, the better the generalization performance, and thus, individual classifiers that correctly classify examples are more important than incorrect ones since the former is helpful to increase the margin of the examples. In addition, we argue that it is more important to increase the margin of examples at the boundary (margin equal to zero), since adding into (or removing from) the ensemble H a classifier would lead to ensembles correctly classifying the examples. Therefore, the proposed measure for ensemble pruning should focus more on correct classifiers and the examples lying near the boundary. Therefore, the importance of individual classifiers can be ordered as shown in Figure 2. Based on the order of importance of individual classifiers, the margin-based measure is proposed in Section 4.

4. Margin-Based Measure

In this section, we propose a heuristic measure for evaluating the importance of individual classifiers based on the gained insight obtained in Section 3: ensemble pruning via directed hill climbing strategy should focus more on examples with small absolute margins as well as classifiers that correctly classify more examples. Several methods use a different approach to calculate diversity during the search.

4.1. Measure for Two-Class Problem. For simplicity of presentation, this section focuses on forward ensemble pruning: given an ensemble subset S which is initialized to be empty, we iteratively add into S the classifier $h \in H \setminus S$. Here, the symbols are similar to the ones in Section 3. Assuming that ensembles use simple majority voting to obtain the predictions, then the margin of an example $\mathbf{x}_i$ of the ensemble S is

$$\text{margin}\left(\mathbf{x}_i\right) = y_i \sum_{t=1}^{|S|} \frac{1}{|S|} h_t\left(\mathbf{x}_i\right) = \frac{y_i \sum_{t=1}^{|S|} h_t\left(\mathbf{x}_i\right)}{|S|}, \tag{9}$$

where $|S|$ is the size of the ensemble S. From (9), $1/|S|$ is the weight of each classifier h_i, $i = 1, 2, \ldots, |S|$, and $y_i h_j(\mathbf{x}_i)/|S|$ is the margin contribution of h_j on the example $\mathbf{x}_i$. Then the proposed measure, margin-based measure (MM), of classifier h with respect to ensemble S and the pruning set D_{pr} is defined as

$$\text{MM}\left(h, S, D_{\text{pr}}\right) = \frac{1}{\left|D_{\text{pr}}\right|} \sum_{\mathbf{x}_i \in D_{\text{pr}}} \text{MM}\left(h, S, \mathbf{x}_i\right), \tag{10}$$

where $\text{MM}(h, S, \mathbf{x}_i)$ is the margin-based measure of h with respect to the subensemble S and current example $\mathbf{x}_i$, defined as

$$\begin{aligned} \text{MM}\left(h, S, \mathbf{x}_i\right) &= \frac{y_i h\left(\mathbf{x}_i\right) / |S|}{\left(\left|\text{margin}\left(\mathbf{x}_i\right)\right| + 1/|S|\right)} \\ &= \frac{y_i h\left(\mathbf{x}_i\right)}{\left(\left|y_i \sum_{j=1}^{|S|} h_j\left(\mathbf{x}_i\right)\right| + 1\right)}, \end{aligned} \tag{11}$$

where the constant parameter $1/|S|$ is to avoid the denominator equal to zero. Since $|y_i \sum h_j(\mathbf{x}_i)| \geq 0$, then $-1 \leq \text{MM}(h, S, \mathbf{x}_i) \leq 1$ and therefore $-1 \leq \text{MM}(h, S, D_{\text{pr}}) \leq 1$. From (9) and (10), $y_i h(\mathbf{x}_i)/|S|$ is exact the margin contribution of h on the example $\mathbf{x}_i$ and $1/|\text{margin}(\mathbf{x}_i)|$ is the weight of h. The rationale of the proposed measure is as follows:

(i) If individual classifier h correctly classifies the example $\mathbf{x}_i$, h increases the margin of $\mathbf{x}_i$, and the corresponding increase value is

$$\frac{y_i h\left(\mathbf{x}_i\right)}{|S|} = \frac{1}{|S|} \tag{12}$$

and thus h favor S correctly classifying $\mathbf{x}_i$, namely, $\text{MM}(h, S, \mathbf{x}_i) \geq 0$ (refer to (10)). If h incorrectly classifies $\mathbf{x}_i$, the prediction of h reduces the margin of $\mathbf{x}_i$ and the reduction is exact

$$\frac{y_i h\left(\mathbf{x}_i\right)}{|S|} = -\frac{1}{|S|} \tag{13}$$

and thus h is harmful to S correctly classifying $\mathbf{x}_i$, namely, $\text{MM}(h, S, \mathbf{x}_i) \leq 0$ (refer to (10)).

(ii) From the discussion of Section 3, $|\text{margin}(\mathbf{x}_i)|$ reflects the confidence that S correctly (or incorrectly) classifies the example $\mathbf{x}_i$. If $|\text{margin}(\mathbf{x}_i)|$ is very small (equal to 0, e.g.), namely, S correctly (or incorrectly) classifying $\mathbf{x}_i$ with a low confidence, adding into S the classifier h may change the prediction of S on the example and therefore h's weight $1/|\text{margin}(\mathbf{x}_i)|$ is large. On the other hand, if $|\text{margin}(\mathbf{x}_i)|$ is very large (equal to 1, e.g.), namely, S correctly (or incorrectly) classifying $\mathbf{x}_i$ with a high confidence, adding into S the classifier h cannot change the prediction of S on the example and therefore h's weight $1/|\text{margin}(\mathbf{x}_i)|$ is small.

The time complexity of calculating (10) or (16) is $O(|S|N)$, which can be $O(N)$ by incrementally updating margins of examples each time a classifier is added to/removed from it, where N is the number of pruning sets. Therefore, the time complexity of ensemble pruning via directed hill climbing

strategy based on the proposed measure is not more than $O(MN)$, where M is the size of the original ensemble learned from training sets.

In this way, the proposed measure focuses more on correct classifiers and the examples lying near the boundaries, which coincides with the conclusions in Section 3.

4.2. Measure for Multiclass Problem. For multiclass classification problem, (11) should be extended so that the proposed measure defined by (10) can deal with the problem.

Let each member h of S map an example $\mathbf{x}_i$ to a label y; namely, $h : \mathbf{x}_i \to y \in [1, L]$, and let

$$V = \left\{ v^{(\mathbf{x}_1)}, \ldots, v^{(\mathbf{x}_N)} \mid v^{(\mathbf{x}_i)} = \left[v_1^{(\mathbf{x}_i)}, \ldots, v_L^{(\mathbf{x}_i)} \right], \ i \in 1, \ldots, N \right\}, \tag{14}$$

where

$v_j^{(\mathbf{x}_i)}$ is the number of votes on the jth label of example $\mathbf{x}_i$ of an ensemble combined by majority voting;

$v_{\max}^{(\mathbf{x}_i)}$ is the number of majority votes on the example $\mathbf{x}_i$;

$v_{\text{sec}}^{(\mathbf{x}_i)}$ is the second largest votes on the example $\mathbf{x}_i$;

$v_{h(\mathbf{x}_i)}^{(\mathbf{x}_i)}$ is the number of votes on label $h(\mathbf{x}_i)$.

From [28], for multiclass, the margin of an example is defined as the difference between the number of correct votes and the maximum number of votes received by any incorrect label; namely,

$$\text{margin}\left(\mathbf{x}_i\right) = \frac{1}{|S|} \left[I\left(S\left(\mathbf{x}_i\right) = y_i\right) \left(v_{y_i}^{(\mathbf{x}_i)} - v_{\text{sec}}^{(\mathbf{x}_i)} \right) - I\left(S\left(\mathbf{x}_i\right) \neq y_i\right) \left(v_{\max}^{(\mathbf{x}_i)} - v_{y_i}^{(\mathbf{x}_i)} \right) \right]. \tag{15}$$

Combining (11) and (15) results in

$$\text{MM}\left(h, S, \mathbf{x}_i\right) = \frac{I\left(\mathbf{x}_i \in e_{tt}\right) - I\left(\mathbf{x}_i \in e_{ft}\right)}{v_{y_i}^{(\mathbf{x}_i)} - v_{\text{sec}}^{(\mathbf{x}_i)} + 1} + \frac{I\left(\mathbf{x}_i \in e_{tf}\right) - I\left(\mathbf{x}_i \in e_{ff}\right)}{v_{\max}^{(\mathbf{x}_i)} - v_{y_i}^{(\mathbf{x}_i)} + 1}, \tag{16}$$

where e_{tt} (or e_{ft}) is the set of examples that are correctly (or incorrectly) classified by current classifier h and correctly classified by the ensemble; similarly e_{tf} (or e_{ff}) is the set of examples that are correctly (or incorrectly) classified by h and incorrectly classified by S. Formally,

$$\begin{aligned} e_{tf} &= \left\{ \mathbf{x}_i \mathbf{x}_i \in D_{\text{pr}} \wedge h\left(\mathbf{x}_i\right) = y_i \wedge S\left(\mathbf{x}_i\right) \neq y_i \right\}, \\ e_{tt} &= \left\{ \mathbf{x}_i \mathbf{x}_i \in D_{\text{pr}} \wedge h\left(\mathbf{x}_i\right) = y_i \wedge S\left(\mathbf{x}_i\right) = y_i \right\}, \\ e_{ft} &= \left\{ \mathbf{x}_i \mathbf{x}_i \in D_{\text{pr}} \wedge h\left(\mathbf{x}_i\right) \neq y_i \wedge S\left(\mathbf{x}_i\right) = y_i \right\}, \\ e_{ff} &= \left\{ \mathbf{x}_i \mathbf{x}_i \in D_{\text{pr}} \wedge h\left(\mathbf{x}_i\right) \neq y_i \wedge S\left(\mathbf{x}_i\right) \neq y_i \right\}. \end{aligned} \tag{17}$$

Table 1: Characteristics of the 18 data sets used in experiments.

ID	Data set	#Insts	#Cls	#Attrs
ID 1	Anneal	898	6	38
ID 2	Audiology	205	7	25
ID 3	Autos	205	7	25
ID 4	Balance-scale	625	3	4
ID 5	Car	1728	6	4
ID 6	Ecoli	336	8	7
ID 7	Flags	194	8	30
ID 8	Glass	214	7	9
ID 9	Horse-colic	368	2	23
ID 10	Hypothyroid	3772	4	29
ID 11	Irish	500	3	6
ID 12	kr-vs-kp	3196	2	37
ID 13	Labor	57	2	16
ID 14	Page-blocks	5473	5	10
ID 15	Segment	2310	7	19
ID 16	Sick	3772	2	30
ID 17	Sonar	208	2	60
ID 18	Wine	178	3	13

In this way, $\text{MM}(h, S, \mathbf{x}_i)$ and thus $\text{MM}(h, S, D_{\text{pr}})$ (the proposed measure) focus more on correct classifiers and the examples lying near the boundaries, which coincide with the conclusions in Section 3.

4.3. Discussion. Unlike other measures where each classifier is independently evaluated, the proposed margin-based measure uses a more global evaluation. Indeed, this criterion involves instance margin values that result from a majority voting of the whole ensemble. Thus, the proposed measure is not only based on individual properties of ensemble members (e.g., accuracy of individual learners). It also takes into account some form of complementarity of classifiers.

From (11), our margin-based measure considers both the correctness of predictions of current classifier and the confidence of prediction of ensemble. Therefore, this measure deliberately favors classifiers with a better performance in classifying low margin samples. Thus, it is a boosting-like strategy which aims to increase the performance on low margin instances. So our strategy of selection will lead to a subset of classifiers with a potentially improved capability to classify complex data in general and border data in particular. Consequently, it will induce a selection of a subset of learners that are designed to efficiently handle minor classes.

From (16), our measure considers the diversity of between ensemble members. Therefore, the measure considers not only the correctness of classifiers, but also the diversity of ensemble members. Therefore, using the proposed measure to prune an ensemble leads to significantly better accuracy results.

5. Experiments

This section first introduces the experiment setting and the characteristics of the data sets used in this paper and

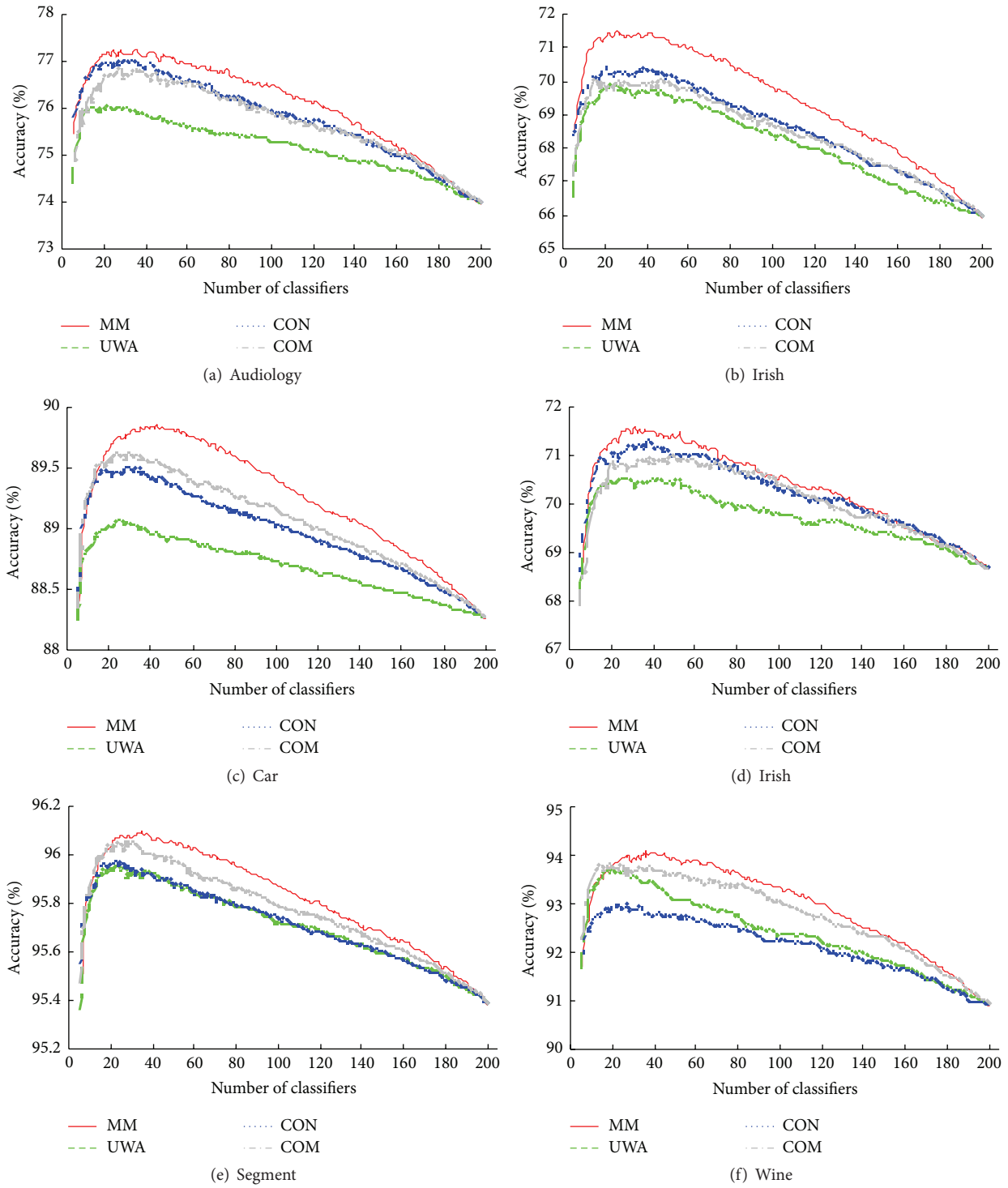

Figure 3: Comparative results for six data sets in the first case.

then reports the comparison of measures for guild ensemble pruning.

5.1. Data Sets and Experimental Setup. We randomly selected 18 data sets from the UCI repository [33]. Each data set was randomly divided into three subsets of equal sizes: one of the subsets as the training set, one as the testing set, and the other as the pruning set. Therefore, we conducted six trials for each data set. We repeated the experiments 50 times and thus conducted a total of 300 trials on each data set. The details of these data sets are summarized in Table 1, where #insts, #Attrs, and #Cls are the size, attribute number, and class number of the corresponding data sets, respectively.

We evaluated the performance of the proposed measure margin-based measure (MM) using forward ensemble selection, where complementariness (COM) [14],

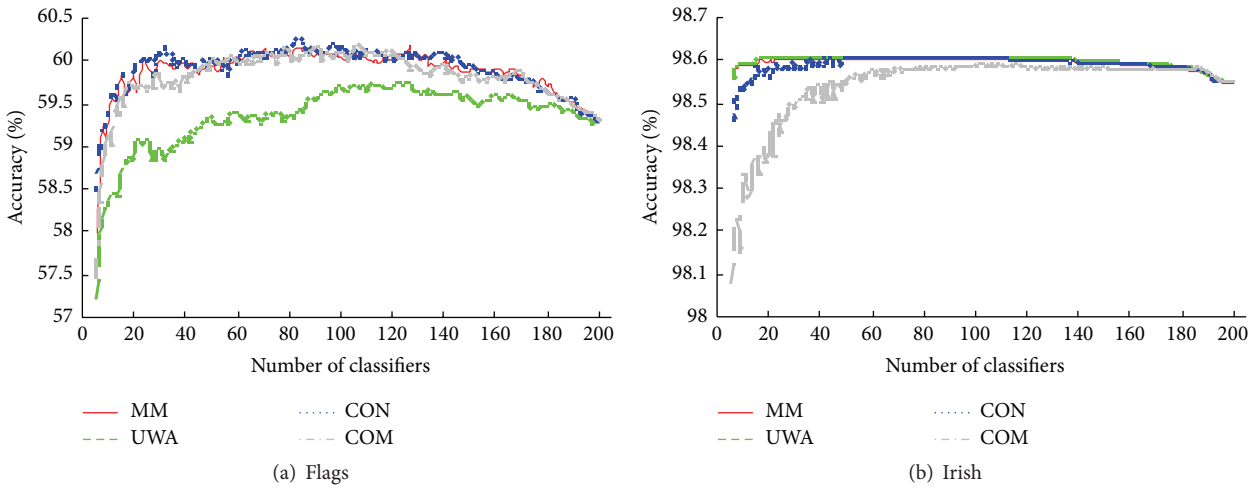

Figure 4: Comparative results for two data sets in the second case.

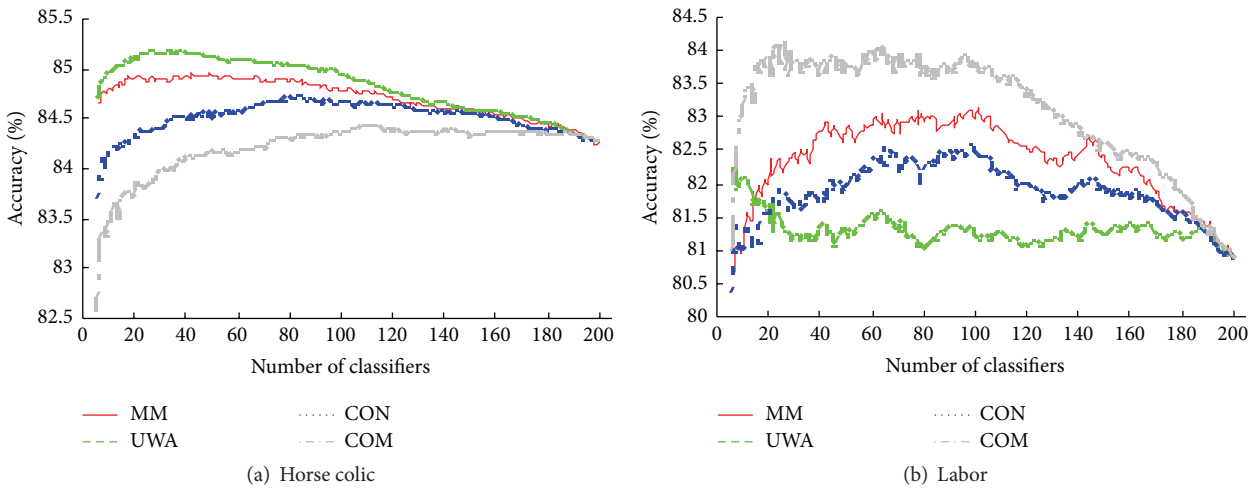

Figure 5: Comparative results for two data sets in the third case.

concurrency (CON) [15], and uncertainty weighted accuracy (UWA) [18] were used as the compared measures. In each trial, a bagging [3] with 200 base classifiers was trained, where the base classifier was J48, which is a Java implementation of C4.5 [34] from Weka [35]. For simplicity, we denote MM, COM, CON, and UMA as the corresponding pruning algorithms supervised by these measures, respectively.

5.2. Accuracy Performance versus the Size of Subensemble. The goal of this experiment was to evaluate the performance of MM by comparing it with UWA, CON, and COM. The experimental results of the 18 tested data sets can be classified into three cases: (1) MM outperforms UWA, CON, and COM; (2) MM performs comparable to one or more of them and outperforms others; and (3) MM is outperformed by one or more of them. The first case contains 13 data sets, the second case contains two data sets, and the last case contains three. Figures 3, 4, and 5 show the representative results from the three cases.

Figure 3 reports the accuracy curves of the four compared measure for six representative data sets that fall into the first case. Results in the figure are reported as average accuracy curves with regard to the number of classifiers, where the horizontal axis is the size of subensembles growing gradually from 5 to 200 with step 1 and the vertical axis is the average accuracy over 300 trials. For the purpose of clarity, the standard deviations are not shown in the figure. The accuracy curves for data sets "audiology," "autos," "car," "glass," "segment," and "wine" are reported in Figures 3(a), 3(b), 3(c), 3(d), 3(e), and 3(f), respectively. Figure 3(a) shows that, with the increase of the number of aggregated classifiers, the accuracy curves of subensembles selected by MM, UWA, CON, and COM increase rapidly, reach the maximum accuracy in the intermediate steps of aggregation which are higher than the accuracy of the whole original ensemble, and then drop until the accuracy is the same as the whole ensemble. The remaining five data sets, "autos," "car," "glass," "segment," and "wine" (shown in Figures 3(b), 3(c), 3(d), 3(e), and 3(f), resp.) have similar accuracy curves to "audiology."

5.3. Summary of Experimental Results. Table 2 summarizes the accuracy of the 300 trials for each data set, where the value in each parentheses is the rank of compared method and the

Table 2: The mean accuracy and ranking of MM, UWA, CON, COM, and bagging. Here, the forward selection is used to pruning ensemble.

Data set	MM	UWA	CON	COM	Bagging
ID 1	98.43 (1.0)	98.32 (2.0)	98.21 (3.0)	98.06 (4.0)	97.34 (5.0)
ID 2	77.17 (1.0)	75.99 (4.0)	77.02 (2.0)	76.74 (3.0)	73.94 (5.0)
ID 3	71.32 (1.0)	69.59 (4.0)	70.28 (2.0)	70.00 (3.0)	65.93 (5.0)
ID 4	84.69 (1.0)	83.92 (4.0)	84.51 (3.0)	84.54 (2.0)	83.32 (5.0)
ID 5	89.79 (1.0)	89.03 (4.0)	89.49 (3.0)	89.62 (2.0)	88.26 (5.0)
ID 6	83.53 (1.0)	83.16 (4.0)	83.35 (3.0)	83.49 (2.0)	82.33 (5.0)
ID 7	59.69 (2.5)	58.95 (5.0)	60.03 (1.0)	59.69 (2.5)	59.29 (4.0)
ID 8	71.43 (1.0)	70.46 (4.0)	71.17 (2.0)	70.76 (3.0)	68.67 (5.0)
ID 9	84.89 (2.0)	85.13 (1.0)	84.40 (3.0)	83.99 (5.0)	84.26 (4.0)
ID 10	99.45 (1.0)	99.43 (2.5)	99.43 (2.5)	99.42 (4.0)	99.29 (5.0)
ID 11	98.60 (1.5)	98.60 (1.5)	98.58 (3.0)	98.48 (5.0)	98.55 (4.0)
ID 12	99.14 (1.0)	99.11 (2.0)	99.05 (3.0)	98.92 (4.0)	98.61 (5.0)
ID 13	82.46 (2.0)	81.19 (4.0)	81.63 (3.0)	83.81 (1.0)	80.84 (5.0)
ID 14	97.19 (1.0)	97.17 (2.0)	97.15 (3.0)	97.11 (4.0)	96.98 (5.0)
ID 15	96.07 (1.0)	95.91 (4.0)	95.93 (3.0)	96.04 (2.0)	95.38 (5.0)
ID 16	98.43 (1.0)	98.40 (2.0)	98.38 (3.0)	98.37 (4.0)	98.20 (5.0)
ID 17	76.02 (3.0)	75.79 (4.0)	76.09 (2.0)	76.45 (1.0)	74.39 (5.0)
ID 18	93.96 (1.0)	93.54 (3.0)	92.83 (4.0)	93.66 (2.0)	90.91 (5.0)
Average rank	1.33	3.17	2.75	2.92	4.83

last row is the average rank. The rank of algorithm is defined as follows: on one data set, the best performing algorithm gets the rank of 1.0, the second best one gets the rank of 2.0, and so on. In the case of ties, average ranks are assigned [36, 37]. The experimental results in Section 5.2 empirically show that MM, UWA, CON, and COM generally reach maximum accuracy when the size of the subensembles is between 20 and 40 (using forward selection for ensemble pruning). Therefore, the subensembles formed by MM with 30 original ensemble members are compared with subensembles formed by UWA, CON, and COM with the same size.

As shown in Table 2, MM outperforms bagging on all the 18 data sets, which indicates that MM efficiently performs ensemble pruning by achieving better predictive accuracies with small subensembles. Table 2 also shows that MM ranks first on 14 out of the 18 data sets and its average rank is 1.33, followed by CON with an average rank of 2.75, COM with an average rank of 2.91, UWA (3.17), and bagging (4.83).

As aforementioned, the backward elimination is another directed hill climbing strategy for ensemble pruning. From experimental results, we observe that performance based on backward elimination strategy is similar to that based on forward selection strategy, and therefore we only present the mean accuracy and ranks of MM, UWA, CON, COM with 30 base classifiers, and bagging (the original ensemble). The corresponding results are illustrated in Table 3. From the table, MM ranks first on 12 data sets and its average rank is 1.42, followed by CON with an average rank of 2.69, COM (2.83), UWA (3.22), and bagging (4.78).

Table 4 shows a summary of the comparisons among the methods, where the pruning methods with "-F" use forward selection to pruning ensemble and similarly, the pruning methods with "-B" use backward elimination to pruning ensemble. The size of each subensemble selected by these ensemble pruning methods is 30. The entry $a_{i,j}$ displays the number of times when the method of the column (j) has a better result than the method of the row (i). The number in the parentheses shows how many of these differences have been statistically significant using pairwise t-tests at the 95% significance level. For example, MM-F has been better than CON-F with pruned trees in 16 of the 18 comparisons and worse in 2. The numbers in the parentheses show that, in 14 cases, the difference in favor of MM-F has been statistically significant; hence, the value in row 3, column 1 of the table is 16 (14).

Table 5 shows the ranking of the comparing methods according to the significant difference between their performances using pairwise t-tests at the 95% significance level. Here, we use all pairwise comparisons as summarized in Table 4. For example, the sum of the numbers in the brackets in the column corresponding to MM-F in Table 4 is 94. The sum of the numbers in the brackets in the row corresponding to MM-F is 10. These are used in Table 5 to calculate the nondominance ranking of MM-F (84).

Tables 4 and 5 demonstrate the significant advantage of MM compared with the best benchmark classifier ensemble methods: CON, COM, and bagging. Besides, compared with ensemble pruning methods using backward elimination, the ones with forward selection show better performance.

6. Conclusion

In this paper, we analysed the importance of individual classifiers with respect to the whole ensembles using margin theory and obtained that ensemble pruning via directed hill climbing strategy should focus more on correct classifiers and the examples lying near the boundary. Based on the derived general principles, we proposed criterion called the

Table 3: The mean accuracy and ranking of MM, UWA, CON, COM, and bagging. Here, the backward elimination is used to pruning ensemble.

Data set	MM	UWA	CON	COM	Bagging
ID 1	98.36 (1.0)	98.27 (2.0)	98.22 (3.0)	98.02 (4.0)	97.34 (5.0)
ID 2	76.88 (2.0)	75.77 (4.0)	77.01 (1.0)	76.35 (3.0)	73.94 (5.0)
ID 3	71.44 (1.0)	69.63 (4.0)	70.39 (2.0)	69.89 (3.0)	65.93 (5.0)
ID 4	84.67 (2.0)	83.93 (4.0)	84.43 (3.0)	84.68 (1.0)	83.32 (5.0)
ID 5	89.89 (1.0)	88.96 (4.0)	89.52 (3.0)	89.60 (2.0)	88.26 (5.0)
ID 6	83.53 (2.0)	83.06 (4.0)	83.35 (3.0)	83.58 (1.0)	82.33 (5.0)
ID 7	60.08 (1.0)	58.82 (5.0)	60.07 (2.0)	59.83 (3.0)	59.29 (4.0)
ID 8	71.36 (1.0)	70.26 (4.0)	71.14 (2.0)	70.67 (3.0)	68.67 (5.0)
ID 9	85.03 (2.0)	85.18 (1.0)	84.38 (5.0)	84.05 (4.0)	84.26 (3.0)
ID 10	99.44 (1.0)	99.42 (3.0)	99.42 (3.0)	99.42 (3.0)	99.29 (5.0)
ID 11	98.60 (1.5)	98.60 (1.5)	98.59 (3.0)	98.45 (5.0)	98.55 (4.0)
ID 12	99.12 (1.0)	99.10 (2.0)	99.06 (3.0)	98.94 (4.0)	98.61 (5.0)
ID 13	81.96 (2.0)	80.96 (4.0)	81.42 (3.0)	83.42 (1.0)	80.84 (5.0)
ID 14	97.19 (1.0)	97.17 (2.0)	97.15 (3.0)	97.11 (4.0)	96.98 (5.0)
ID 15	96.05 (1.0)	95.90 (4.0)	95.93 (3.0)	96.04 (2.0)	95.38 (5.0)
ID 16	98.40 (1.0)	98.39 (2.5)	98.39 (2.5)	98.36 (4.0)	98.20 (5.0)
ID 17	76.12 (3.0)	75.72 (4.0)	76.25 (2.0)	76.65 (1.0)	74.39 (5.0)
ID 18	93.84 (1.0)	93.44 (3.0)	92.89 (4.0)	93.75 (2.0)	90.91 (5.0)
Average rank	1.42	3.22	2.69	2.83	4.78

Table 4: Summary of results.

	MM-F	UWA-F	CON-F	COM-F	MM-B	UWA-B	CON-B	COM-B	Bagging
MM-F	—	1 (1)	2 (0)	2 (1)	7 (4)	2 (2)	2 (0)	4 (2)	0 (0)
UWA-F	16 (14)	—	10 (6)	11 (8)	17 (13)	4 (1)	10 (7)	11 (7)	1 (0)
CON-F	16 (12)	8 (6)	—	7 (4)	17 (11)	7 (5)	11 (0)	7 (6)	0 (0)
COM-F	16 (11)	7 (6)	11 (6)	—	16 (10)	7 (5)	11 (7)	9 (0)	2 (0)
MM-B	11 (4)	1 (0)	1 (0)	2 (1)	—	1 (1)	2 (0)	4 (2)	0 (0)
UWA-B	16 (14)	14 (6)	11 (8)	11 (9)	16 (15)	—	11 (8)	11 (8)	1 (0)
CON-B	16 (12)	8 (5)	7 (0)	7 (4)	16 (10)	7 (4)	—	7 (4)	0 (0)
COM-B	14 (11)	7 (6)	11 (7)	9 (1)	14 (11)	7 (6)	11 (9)	—	2 (0)
Bagging	18 (16)	17 (15)	18 (14)	16 (15)	18 (15)	17 (15)	18 (14)	16 (15)	—

Table 5: Rank of the methods using the significant differences from all pairwise comparisons.

Method	Dominance rank (wins-losses)	Wins	Losses
MM-F	84	94	10
MM-B	81	89	8
CON-B	6	45	39
COM-F	−2	43	45
CON-F	−3	41	44
COM-B	−7	44	51
UWA-F	−11	45	56
UWA-B	−29	39	68
Bagging	−119	0	119

margin-based measure to explicitly evaluate the importance of individual classifiers. Experimental comparisons on 18 UCI data sets showed that the proposed measure outperforms other state-of-the-art measures and the original ensemble.

The proposed metric in this paper can apply not only to ensemble pruning based on directed hill climbing search but also to other ensemble pruning methods. Therefore, more experiments will be conducted to evaluate the performance of the proposed measure.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China (Grants no. 61472370, no. 61202207, no. 61501393, no. 61402393, and no. 61572417),

Project of Science and Technology Department of Henan Province (no. 162102210310, no. 152102210129, and no. 142400410486), and Science and Technology Research key Project of the Education Department of Henan Province (Grant no. 15A520026).

References

[1] X. Wu, V. Kumar, Q. J. Ross et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[2] T. G. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, pp. 1–15, Cagliari, Italy, June 2000.

[3] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[4] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, part 2, pp. 119–139, 1997.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] F. T. Liu, K. M. Ting, Y. Yu, and Z.-H. Zhou, "Spectrum of variable-random trees," *Journal of Artificial Intelligence Research*, vol. 32, no. 1, pp. 355–384, 2008.

[7] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[8] D. Zhang, S. Chen, Z. Zhou, and Q. Yang, "Constraint projections for ensemble learning," in *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI '08)*, pp. 758–763, Chicago, Ill, USA, July 2008.

[9] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[10] L. Guo, *Margin framework for ensemble classifiers. Application to remote sensing data [Ph.D. thesis]*, University of Bordeaux 3, Pessac, France, 2011.

[11] Z. Ma, Q. Dai, and N. Liu, "Several novel evaluation measures for rank-based ensemble pruning with applications to time series prediction," *Expert Systems with Applications*, vol. 42, no. 1, pp. 280–292, 2015.

[12] W. M. Zhi, H. P. Guo, M. Fan, and Y. D. Ye, "Instance-based ensemble pruning for imbalanced learning," *Intelligent Data Analysis*, vol. 19, no. 4, pp. 779–794, 2015.

[13] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[14] G. Martinez-Muverbnoz and A. Suarez, "Aggregation ordering in bagging," in *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, pp. 258–263, Acta Press, Calgary, Canada, 2004.

[15] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "Ensemble diversity measures and their application to thinning," *Information Fusion*, vol. 6, no. 1, pp. 49–62, 2005.

[16] I. Partalas, G. Tsoumakas, and I. P. Vlahavas, "Focused ensemble selection: a diversity-based method for greedy ensemble selection," in *Proceedings of the 18th European Conference on Artificial Intelligence*, pp. 117–121, Patras, Greece, July 2008.

[17] G. Martinez-Muñoz, D. Hernández-Lobato, and A. Suarez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 245–259, 2009.

[18] I. Partalas, G. Tsoumakas, and I. P. Vlahavas, "An ensemble uncertainty aware measure for directed hill climbing ensemble pruning," *Machine Learning*, vol. 81, no. 3, pp. 257–282, 2010.

[19] Z. Lu, X. D. Wu, X. Q. Zhu, and J. Bongard, "Ensemble pruning via individual contribution ordering," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 871–880, Washington, DC, USA, July 2010.

[20] L. Guo and S. Boukir, "Margin-based ordered aggregation for ensemble pruning," *Pattern Recognition Letters*, vol. 34, no. 6, pp. 603–609, 2013.

[21] C. Qian, Y. Yu, and Z. H. Zhou, "Pareto ensemble pruning," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2935–2941, Austin, Tex, USA, January 2015.

[22] B. Krawczyk and M. Woźniak, "Untrained weighted classifier combination with embedded ensemble pruning," *Neurocomputing*, vol. 196, pp. 14–22, 2016.

[23] Y. Zhang, S. Burer, and W. N. Street, "Ensemble pruning via semi-definite programming," *Journal of Machine Learning Research*, vol. 7, pp. 1315–1338, 2006.

[24] W. M. Zhi, H. P. Guo, and M. Fan, "Energy-based metric for ensemble selection," in *Proceedings of the 14th Asia-Pacific Web Conference*, pp. 306–317, Kunming, China, April 2012.

[25] W. Fan, F. Chu, H. Wang, and P. S. Yu, "Pruning and dynamic scheduling of cost-sensitive ensembles," in *Proceedings of the 18th National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, Edmonton, Canada, August 2002.

[26] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 137–144, Banff, Canada, July 2004.

[27] Q. Dai and M. L. Li, "Introducing randomness into greedy ensemble pruning algorithms," *Applied Intelligence*, vol. 42, no. 3, pp. 406–429, 2015.

[28] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[29] I. Partalas, G. Tsoumakas, and I. Vlahavas, "A study on greedy algorithms for ensemble pruning," Tech. Rep. TR-LPIS-360-12, LPIS, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece, 2012.

[30] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 211–218, Nashville, Tenn, USA, September 1997.

[31] Q. Dai, T. Zhang, and N. Liu, "A new reverse reduce-error ensemble pruning algorithm," *Applied Soft Computing*, vol. 28, pp. 237–249, 2015.

[32] W. Gao and Z.-H. Zhou, "On the doubt about margin explanation of boosting," *Artificial Intelligence*, vol. 203, pp. 1–18, 2013.

[33] A. Asuncion and D. Newman, "UCI Machine Learning Repository," 2007.

[34] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, Calif, USA, 1993.

[35] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[36] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[37] S. García and F. Herrera, "An extension on 'statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.