Diversity Regularized Ensemble Pruning

Nan Li[1,2], Yang Yu[1], and Zhi-Hua Zhou[1]

[1] National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210046, China
[2] School of Mathematical Sciences, Soochow University, Suzhou 215006, China
{lin, yuy, zhouzh}@lamda.nju.edu.cn

Abstract. Diversity among individual classifiers is recognized to play a key role in ensemble, however, few theoretical properties are known for classification. In this paper, by focusing on the popular ensemble pruning setting (i.e., combining classifier by *voting* and measuring diversity in *pairwise* manner), we present a theoretical study on the effect of diversity on the generalization performance of voting in the PAC-learning framework. It is disclosed that the diversity is closely-related to the hypothesis space complexity, and encouraging diversity can be regarded to apply regularization on ensemble methods. Guided by this analysis, we apply explicit diversity regularization to ensemble pruning, and propose the *Diversity Regularized Ensemble Pruning* (DREP) method. Experimental results show the effectiveness of DREP.

Key words: diversity, ensemble pruning, diversity regularization

1 Introduction

Ensemble methods [33], which train multiple classifiers for one single task, are among the state-of-the-art machine learning approaches. It is widely accepted that ensemble methods usually achieve better generalization performance than single classifiers, and they have achieved great successes in a large variety of real-word applications.

Generally speaking, an ensemble is built in two steps: first multiple classifiers are trained for one task, and then these classifiers are combined together to get a better performance in some manners like voting. Given multiple trained individual classifiers, instead of combining all of them, there are many studies tring to select a subset from them to comprise the ensemble [28]. In the literature, the task of reducing ensemble sizes is called as ensemble pruning [19], selective ensemble [35], ensemble selection [6] or ensemble thinning [1]. Currently, we do not distinguish between them and use ensemble pruning for simplicity. By producing ensembles of smaller sizes, ensemble pruning has the apparent advantage of improving storage and computational efficiency for predictions. Furthermore, both theoretical and empirical studies have shown that ensemble pruning can also improve the generalization performance of ensemble [35, 6, 32, 20], that is, the pruned ensemble can achieve better performance than the complete ensemble.

In the ensemble pruning literature, greedy pruning methods which search the space of possible classifier subsets by taking greedy local search have drawn much attention [24, 20], it is because compared with methods that directly select optimal or near-optimal classifier subsets, they are able to achieve comparative performance and robustness at much smaller computational costs. There are two salient parameters in greedy pruning methods: the direction for searching the space (i.e., forward and backward) and the criterion for evaluating available actions at each search step. Since it is shown that the direction does not significantly affect the performance [24], much attention has been paid on the design of the evaluation criterion. As diversity among individual classifiers is widely recognized to be key to the success of an ensemble, many evaluation criteria have been developed to select diverse individual classifiers, mainly by smart heuristics [1, 23, 20, 24]. In practice, although positive correlation has been demonstrated between diversity and accuracy of ensemble [9, 16, 11], few theoretical prosperities of ensemble diversity is known. Moreover, the usefulness of exploiting diversity measures in building stronger ensemble was doubted in [15, 27].

In this paper, concentrating on a popular setting of ensemble pruning where the individual classifiers are combined by *voting* and the diversity is measured in the *pairwise* manner, we present a theoretical analysis on the effect of diversity on the generalization performance of voting based on the probably approximately correct learning (PAC-learning) framework [29]. To our best knowledge, this is the first PAC-style analysis on diversity's effect on voting. We show that encouraging larger diversity leads to smaller hypothesis space complexity and thus better generalization performance, which implies that controlling diversity can be regarded to apply regularization on ensemble methods. Then, guided by the theoretical analysis, we propose the DREP method which is a greedy forward ensemble pruning method with explicit diversity regularization. Experimental results show the effectiveness of the DREP method.

The remainder of the paper is organized as follows. Section 2 gives a brief review on ensemble selection and ensemble diversity. Section 3 presents our theoretical study on the role of diversity in voting. Based on the theoretical results, Section 4 proposes the DREP method, followed by Section 5 which reports on the experimental results. Finally, the paper is concluded in Section 6.

2 Related Work

With the goal of improving storage and computational efficiency as well as generalization performance, ensemble pruning deals with the problem of reducing ensemble sizes. The first work on this topic was possibly done by Margineantu and Dietterich [19], which tried to prune AdaBoost, but later Tamon and Xiang [26] showed that the boosting pruning problem is intractable even to approximate. Instead of pruning ensembles generated by sequential methods, Zhou et al. [35] and Caruana et al. [6] respectively studied on pruning ensembles generated by parallel methods such as Bagging [3] and parallel heterogeneous ensembles consisting of different types of individual classifiers, and it was shown that better

performance can be obtained at smaller ensemble sizes. Ever since, most ensemble pruning studies were devoted to parallel ensemble methods.

Given a set of trained classifiers, selecting the sub-ensemble with the best generalization performance is difficult mainly due to two reasons: First, it is not easy to estimate the generalization performance of a sub-ensemble; second, finding the optimal subset is a combinatorial search problem with exponential computational complexity, thus it is unfeasible to compute the exact solution by exhaustive search and approximate search is needed. In the past decade, a number of methods have been proposed to overcome this difficulty [28], which can be roughly classified into two groups based on their employed search methods. The first group of methods use *global search* to directly select the optimal or near-optimal classifier subset. In the literature, many techniques have been used, such as genetic algorithm [35], semi-definite programming [31], clustering [12, 17], sparse optimization with sparsity-inducing prior [7] or ℓ_1-norm constraint [18], etc. In practice, this kind of methods can achieve good performance, but their computational costs are usually quite large.

The second group of ensemble pruning methods is based on *greedy local search* of the space of all possible ensemble subsets [20, 24]. According to the search direction, this group of methods can be further divided into greedy *forward* pruning methods which start with empty set and iteratively add the classifier optimizing certain criterion, and greedy *backward* methods that start with the complete ensemble and iteratively eliminate classifiers. It has been shown that greedy pruning methods are able to achieve comparative performance and robustness with global search methods but at much smaller computational costs [20, 13]. Moreover, based on extensive experiments, Partalas et al. [24] suggested to use the greedy forward methods because both directions achieve similar performance but the forward direction produces smaller ensemble sizes. Then, the study of greedy pruning methods was mainly devoted to the criterion that is used for evaluating available actions at each local search step. Since the diversity within an ensemble is widely recognized to be important to its success, many criteria have been proposed to select diverse individual classifiers, such as Kappa [19, 1], complementarity [21, 20], orientation [22, 20], margin distance [21, 20], FES [23], etc. It is easy to see that most of these criteria are based on smart heuristics.

In practice, the importance of diversity was first discovered from error analysis for regression [14], and then extended to classification. For classification, it has been observed from empirical studies like [9] that there exists positive correlation between diversity and accuracy of ensemble. Also, some theoretical studies have shown that encouraging diversity is beneficial. For example, Kuncheva et al. [16] found that negative dependence between individual classifiers is beneficial to the accuracy of an ensemble, Fumera and Roli [11] found that the performance of ensemble depends on the performance of individual classifiers and their correlation. Based on current results, we can see that it is no problem to reach that encouraging diversity is beneficial to the performance of ensemble, but it is hard to tell the theoretical properties of diversity in ensemble. In the famous margin explanation of voting [25], the diversity is totally not considered in the framework.

Also, some doubts have been raised on the usefulness of exploiting diversity measures in building stronger ensembles [15, 27]. Therefore, understanding ensemble diversity remains an important issue in ensemble learning and further investigations are needed. Recently, by defining diversity and ensemble combination rule in the parameter space, our previous work [30] showed that diversity control can play a role of regularization as in statistical learning methods, which, however, relies on linear classifiers and average combination and thus cannot be applied to other kind of classifiers and voting combination. In this work, we consider the popular ensemble pruning setting, i.e., the diversity is measured in the output space which leaves the specification of classifiers unimportant, and the individual classifiers are combined by voting.

3 Diversity and Generalization Performance of Voting

In ensemble pruning, voting is one of the most widely used methods to combine individual classifiers. In this section, we give a theoretical study on the effect of diversity on the generalization performance of voting.

3.1 Basics and Diversity Measure

Consider binary classification, given a set of n trained classifiers $H = \{h_i(\boldsymbol{x})\}_{i=1}^n$, where each classifier $h_i : \mathcal{X} \mapsto \{-1, +1\}$ is a mapping from the feature space $\mathcal{X}$ to the class label set $\{-1, +1\}$, the voting rule defines a decision function by taking an average of classifiers in H as

$$f(\boldsymbol{x}; H) = \frac{1}{n} \sum\nolimits_{i=1}^{n} h_i(\boldsymbol{x}) \ , \tag{1}$$

and it predicts the class label of $\boldsymbol{x}$ as $\mathtt{sign}[f(\boldsymbol{x}; H)]$. Obviously, it makes wrong prediction on example $(\boldsymbol{x}, y)$ only if $yf(\boldsymbol{x}; H) \leq 0$, and $yf(\boldsymbol{x}; H)$ is called the *margin* of f at $(\boldsymbol{x}, y)$.

Let $\mathcal{D}$ is the underlying distribution over $\mathcal{X} \times \{-1, +1\}$, and $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ is a set of examples randomly sampled from $\mathcal{D}$, the *generalization error* (denoted as $err_g(f)$) and the *empirical error* with margin θ on S (denoted as $err_S^\theta(f)$) are respectively defined as

$$err_g(f) = P_{(\boldsymbol{x},y)\sim\mathcal{D}}[yf(\boldsymbol{x}) \leq 0] \quad \text{and} \quad err_S^\theta(f) = \frac{1}{m} \sum\nolimits_{i=1}^{m} \mathbb{I}[y_i f(\boldsymbol{x}_i) \leq \theta] \ , \tag{2}$$

where $\mathbb{I}[z]$ is the indicator function which takes 1 if z is `true`, and 0 otherwise.

Although there is no generally accepted formal definition of diversity in the literature [5, 34], popular diversity measures are usually formalized based on pairwise difference between every pair of individual classifiers [15], such as Q-statistics, correlation coefficient, disagreement measure and κ-statistics. In this work, we also measure diversity based on pairwise difference, and the definition is given as follows.

Definition 1 *Given a set of m examples $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, the diversity of classifier set $H = \{h_i(\boldsymbol{x})\}_{i=1}^n$ on S is defined as*

$$\texttt{div}(H) = 1 - \frac{1}{\sum_{1 \le i \ne j \le n} 1} \sum\nolimits_{1 \le i \ne j \le n} \texttt{diff}(h_i, h_j) \ , \tag{3}$$

where $\texttt{diff}(\cdot, \cdot)$ measures the pairwise difference between two classifiers as

$$\texttt{diff}(h_i, h_j) = \frac{1}{m} \sum\nolimits_{k=1}^m h_i(\boldsymbol{x}_k) h_j(\boldsymbol{x}_k) \ . \tag{4}$$

It is obvious that the difference $\texttt{diff}(h_i, h_j)$ falls into the interval $[-1, 1]$, and $\texttt{diff}(h_i, h_j)$ equals to 1 (or -1) only if two classifiers h_i and h_j always make the same (or opposite) predictions on the data sample S, and the smaller $\texttt{diff}(h_i, h_j)$, the larger difference between h_i and h_j. Consequently, since the diversity is based on the average of pairwise differences, we can see that the larger $\texttt{div}(H)$ the larger the diversity of the classifier set H.

It is easy to find that this diversity measure is closely-related with the disagreement measure [15]. Moreover, different from [30] which defines the diversity in the parameter space of classifiers, here this diversity measure is defined in the output space, thus can cover various kinds of individual classifiers.

3.2 Theoretical Results

Our analysis is based on the PAC-learning framework [29], which is one of the most widely used framework for analyzing learning algorithms. Before giving the main results, we first introduce some necessary background.

In learning theory, it is known that the generalization error of a learning algorithm can be bounded by its empirical error and the complexity of feasible hypothesis space [29, 2]. Since the hypothesis space is uncountable for many learning methods, the hypothesis space complexity is often described by a quantity called *covering number*, which is defined as below.

Definition 2 *Let B be a metric space with metric ρ. Given a set of m examples $S = \{\boldsymbol{x}_i\}_{i=1}^m$ and a function space $\mathcal{F}$, characterize every $f \in \mathcal{F}$ with a vector $\boldsymbol{v}_S(f) = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]^\top \in B^m$. The covering number in p-norm $\mathcal{N}_p(\mathcal{F}, \epsilon, S)$ is the minimum number l of vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in B^m$ such that, for all $f \in \mathcal{F}$ there exists $j \in \{1, \ldots, l\}$,*

$$\|\rho(\mathbf{v}_S(f), \mathbf{u}_j)\|_p = \left(\sum\nolimits_{i=1}^m \rho(f(\boldsymbol{x}_i), u_{j,i})^p \right)^{1/p} \le m^{1/p} \epsilon \ ,$$

and $\mathcal{N}_p(\mathcal{F}, \epsilon, m) = \sup_{S:|S|=m} \mathcal{N}_p(\mathcal{F}, \epsilon, S)$.

Currently, we show how the ensemble diversity affects the generalization performance of voting. In particular, our study mainly focuses on the effect of diversity on the hypothesis space complexity of voting. Before presenting the main result, we give the following lemma.

Lemma 1 *Given a set of classifiers* $H = \{h_i(\boldsymbol{x})\}_{i=1}^n$ *and a set of examples* $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$*, denote* $\mathbf{f} = [f(\boldsymbol{x}_1; H), \ldots, f(\boldsymbol{x}_m; H)]^\top$ *be the output of the decision function* f*'s outputs on* S*. On data set* S*, if* $\mathtt{div}(H) \geq q$*, then it follows*

$$\|\mathbf{f}\|_1 \leq m\sqrt{1/n + (1-1/n)(1-q)} \ .$$

Proof. By basic algebra, we have

$$\begin{aligned}\|\mathbf{f}\|_2^2 &= \sum\nolimits_{i=1}^m \left(\frac{1}{n}\sum\nolimits_{t=1}^n h_t(\boldsymbol{x}_i)\right)^2 = \sum\nolimits_{i=1}^m \left(\frac{1}{n} + \frac{1}{n^2}\sum\nolimits_{1\leq j\neq k\leq n} h_j(\boldsymbol{x}_i)h_k(\boldsymbol{x}_i)\right) \\ &= m\left(1/n + (1-\mathtt{div}(H))(1-1/n)\right) \geq 0 \ .\end{aligned}$$

We can find the quantity $1/n + (1-q)(1-1/n)$ is always non-negative. Then, based on the inequality $\|\mathbf{f}\|_1 \leq \sqrt{m}\|\mathbf{f}\|_2$, we can obtain the result directly. □

Theorem 1. *Let* $\mathcal{F}$ *denote the function space such that for every* $f \in \mathcal{F}$*, there exist a set of* n *classifiers* $H = \{h_i(\boldsymbol{x})\}_{i=1}^n$ *satisfying* $f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n h_i(\boldsymbol{x})$ *and* $\mathtt{div}(H) \geq q$ *for any i.i.d. sample* S *of size* m*, then for any* ϵ*, it holds*

$$\log_2 \mathcal{N}_\infty(\mathcal{F}, \epsilon, m) \leq \ \frac{36(1+\ln n)}{\epsilon^2}\log_2\left(2m\lceil 4\sqrt{1/n + (1-1/n)(1-q)}/\epsilon + 2\rceil + 1\right) .$$

Proof. This proof follows similar strategy with Theorem 4 and 5 in [31], here we give the main sketch and focus on the difference. If $\epsilon \geq 1$, the result follows trivially, so it is assumed $\epsilon \leq 1$ subsequently. First, the interval $[-1-\epsilon/2, 1+\epsilon/2]$ is divided into $n = \lceil 4/\epsilon + 2\rceil$ sub-intervals, each of size no larger than $\epsilon/2$, and θ_j be the boundaries of the sub-intervals so that $\theta_j - \theta_{j-1} \leq \epsilon/2$ for all j. Let $j_l(i)$ denote the maximum index of θ_j such that $f(\boldsymbol{x}_i) - \theta_{j_l(i)} \geq \epsilon/2$ and $j_r(i)$ the maximum index of θ_j such that $f(\boldsymbol{x}_i) - \theta_{j_r(i)} \leq -\epsilon/2$. Let

$$\boldsymbol{h}_i = [h_1(\boldsymbol{x}_i), \ldots, h_T(\boldsymbol{x}_i)]^\top, \quad \boldsymbol{h}'_i = [\boldsymbol{h}_i, -\theta_{j_l(i)}]^\top \quad \text{and} \quad \boldsymbol{h}''_i = [-\boldsymbol{h}_i, \theta_{j_r(i)}]^\top.$$

Then, based on similar steps in [31], the covering number $\mathcal{N}_\infty(\mathcal{F}, \epsilon, S)$ is no more than the number of possible values of the vector $\boldsymbol{\beta}$, which is defined as

$$\boldsymbol{\beta} = g_p\left(\sum\nolimits_{i=1}^m a_i\boldsymbol{h}'_i + \sum\nolimits_{i=1}^m b_i\boldsymbol{h}''_i\right) \ , \tag{5}$$

where $g_p(\boldsymbol{u})$ is a component-wise function mapping each component u_i of $\boldsymbol{u}$ to $p \cdot \mathtt{sign}(u_i)|u_i|^{p-1}$ with $p \geq 2$, and a_i's and b_i's are non-negative integers under the constraint

$$\sum\nolimits_{i=1}^m (a_i + b_i) \leq 36(1+\ln n)/\epsilon^2 \ . \tag{6}$$

It is easy to find that there is an one-to-one mapping between $\boldsymbol{h}'_i$ and $\boldsymbol{h}''_i$, so the number of possible values of $\boldsymbol{h}'_i$ and $\boldsymbol{h}''_i$ equals to that of $\boldsymbol{h}'_i$. Let $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]^\top$ be f's outputs on S, based on Lemma 1, we can obtain that $\|\boldsymbol{f}\|_1 \leq m\sqrt{1/n + (1-1/n)(1-q)}$. Then, based on the definition of $\theta_{j_l(i)}$, we can find that the number of possible values of $\boldsymbol{h}'_i$ is no more than

$$m\lceil 4\sqrt{1/n + (1-1/n)(1-q)}/\epsilon + 2\rceil \ .$$

Consequently, from (5) and (6) we can find that the number of possible values of $(\boldsymbol{\beta}, z)$ is upper-bounded by

$$\left(2m\lceil 4\sqrt{1/n+(1-1/n)(1-q)}/\epsilon+2\rceil+1\right)^{36(1+\ln n)/\epsilon^2} ,$$

which completes the proof. □

Furthermore, based on Theorem 1 we can obtain the relationship between diversity and generalization performance of voting, which is given as follows.

Corollary 1 *Under the assumptions of Theorem 1, with probability at least* $1-\delta$*, for any* $\theta > 0$*, every function* $f \in \mathcal{F}$ *satisfies the following bound*

$$err_g(f) \leq err_S^{\theta}(f) + \frac{C}{\sqrt{m}}\left(\frac{\ln n \ln\left(m\sqrt{1/n+(1-1/n)(1-q)}\right)}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2} ,$$

where C *is a constant.*

Proof. Based on Bartlett's Lemma 4 in [2], we can obtain

$$err_g(f) \leq err_S^{\theta}(f) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}_{\infty}(\mathcal{F}, \epsilon/2, 2m) + \ln\frac{2}{\delta}\right)} . \tag{7}$$

By applying Theorem 1 on (7), we can obtain the result. □

Above results show that, when other factors are fixed, encouraging high diversity among individual classifiers (i.e., large value of q in Theorem 1 and Corollary 1) will make the hypothesis space complexity of voting small, and thus better generalization performance can be expected.

3.3 Remarks and Discussions

It can be observed from above theoretical analysis that the diversity is directly related to the hypothesis space complexity of voting, and then affects its generalization performance. From the view of statistical learning, controlling ensemble diversity has a direct impact on the size of hypothesis space of voting, indicating that it plays a role similar with regularization as in popular statistical learning methods. In other words, it implies that encouraging diversity can be regarded to apply regularization on ensemble methods. Also, this result show that encouraging diversity is beneficial but not straightforwardly related to the ensemble accuracy, which coincides with previous study in [16].

To our best knowledge, this work provides the first PAC-style analysis on the role of diversity in voting. The margin explanation of voting presented in [25] is also in the PAC-learning framework, but it is obvious that our work is significantly different because diversity is considered explicitly. The hypothesis space complexity of voting becomes small when the diversity increases, but it is simply characterized by the VC-dimension of individual classifier in [25]. Intuitively, due to the diversity, some parts of the hypothesis space of voting are infeasible, excluding these parts leads to tighter bounds, while assuming the hypothesis space compact makes the bounds looser.

4 Diversity Regularized Ensemble Pruning

In this section, we apply above theoretical analysis to ensemble pruning, and propose the *Diversity Regularized Ensemble Pruning* (DREP) method, which is a greedy forward ensemble pruning method.

The main difference between DREP and existing greedy pruning methods lies in the criterion for evaluating available actions at each step. In the previous section, it is shown in Corollary 1 that the generalization performance of an ensemble depends on its empirical error and diversity, so it is natural to design the evaluation criterion accordingly. However, it is easy to see that in the bound the diversity has complicated operations with factors including the sample size m, number of classifier n, η and θ, etc. Then for a given problem, it will be difficult to specify the tradeoff between empirical error and diversity. Hence, a tradeoff parameter is involved in the proposed method.

Moreover, it is easy to see from (3) that when we want to evaluate the diversity of a new ensemble which is obtained by adding one individual classifier, it is needed to compute the pairwise difference between the new added classifier and all the existing classifiers. As a consequence, at each step, if there are many candidate individual classifiers, directly evaluating diversity based on the definition will be of high computational complexity. To avoid this issue, we use a more efficient way based on the following proposition.

Propsition 1 *Given a classifier $h'(\boldsymbol{x})$ and a classifier set $H = \{h_i(\boldsymbol{x})\}_{i=1}^n$, let $H' = H \cup \{h'(\boldsymbol{x})\}$, the diversity of H' on $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ is*

$$\texttt{div}(H') = \frac{2}{n+1} + \frac{n-1}{n+1}\,\texttt{div}(H) - \frac{2}{n+1}\,\texttt{diff}(h', H) \tag{8}$$

where $\texttt{div}(H)$ is the diversity of H on S and $\texttt{diff}(h', H)$ measures the difference between new classifier $h'(\boldsymbol{x})$ and H as

$$\texttt{diff}(h', H) = \frac{1}{m}\sum\nolimits_{i=1}^{m} h'(\boldsymbol{x}_i) f(\boldsymbol{x}_i; H)\ . \tag{9}$$

and $f(\boldsymbol{x}; H)$ is the decision function of H defined in (1).

Proof. Based on the definitions in (3) and (4), it is not hard to obtain

$$\begin{aligned}
\texttt{div}(H') &= 1 - \frac{1}{n(n+1)}\left(\sum\nolimits_{1\le i\ne j\le n} \texttt{diff}(h_i, h_j) + 2\sum\nolimits_{i=1}^{n} \texttt{diff}(h', h_i)\right) \\
&= 1 - \frac{1}{n(n+1)}\left(n(n-1)(1-\texttt{div}(H)) + \frac{2}{m}\sum\nolimits_{i=1}^{m}\left(h'(\boldsymbol{x}_i)\sum\nolimits_{k=1}^{n} h_k(\boldsymbol{x}_i)\right)\right) \\
&= \frac{2}{n+1} + \frac{n-1}{n+1}\,\texttt{div}(H) - \frac{2}{m(n+1)}\sum\nolimits_{i=1}^{m} h'(\boldsymbol{x}_i) f(\boldsymbol{x}; H)\ ,
\end{aligned}$$

which leads to the result directly. □

Algorithm 1 The DREP method

Input: ensemble to be pruned $H = \{h_i(\boldsymbol{x})\}_{i=1}^{n}$, validation data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ and tradeoff parameter $\rho \in (0, 1)$

Output: pruned ensemble H^*

1: initialize $H^* \leftarrow \emptyset$
2: $h(\boldsymbol{x}) \leftarrow$ the classifier in H with the lowest error on S
3: $H^* \leftarrow \{h(\boldsymbol{x})\}$ and $H \leftarrow H \setminus \{h(\boldsymbol{x})\}$
4: **repeat**
5: **for** each $h'(\boldsymbol{x})$ in H **do**
6: compute $d_{h'} \leftarrow \texttt{diff}(h', H^*)$ based on (9)
7: **end for**
8: sort classifiers $h'(\boldsymbol{x})'s$ in H in the ascending order of $d_{h'}$'s
9: $\Gamma \leftarrow$ the first $\lceil \rho \cdot |H| \rceil$ classifiers in the sorted list
10: $h(\boldsymbol{x}) \leftarrow$ the classifier in Γ which most reduces the error of H^* on S
11: $H^* \leftarrow \{h(\boldsymbol{x})\}$ and $H \leftarrow H \setminus \{h(\boldsymbol{x})\}$
12: **until** the error of H^* on S cannot be reduced

It can be found that at each step of greedy forward pruning, $\texttt{div}(H)$ is a constant and $\texttt{div}(H')$ is a monotonically decreasing function of $\texttt{diff}(h', H)$. Thus, at each step, the task of estimating diversity $\texttt{div}(H')$ can be substituted by computing the difference $\texttt{diff}(h', H)$. The candidate classifier h' that can achieve smaller $\texttt{diff}(h', H)$ will lead to larger diversity $\texttt{div}(H')$. Obviously, in such a manner, we only need to compute the difference between the candidate classifier and the decision function of existing sub-ensemble rather than each of its members, which reduces the computational cost heavily comparing with the computation of diversity from scratch.

The pseudocode of the DREP method is presented in Algorithm 1. Specifically, it has three inputs: the ensemble H to be pruned, the validate data set S which is used to estimate empirical error and diversity and the tradeoff parameter ρ. Starting with the classifier with lowest error on validation set (lines 2-3), the DREP method iteratively selects classifier based on both empirical error and diversity. Concretely, at each step it first sorts the candidate classifiers in the ascending order of their differences with current sub-ensemble (lines 5-8), and then from the front part of sorted list it selects the classifier which can most reduce the empirical error on the validate data set. It can be found from Proposition 1 that the front classifiers will lead to large ensemble diversity. Also, among the front classifiers the one which reduces the empirical error most will be selected, thus it can be expected that the obtained ensemble will have both large diversity and small empirical error. These two criteria are balanced by the parameter ρ, i.e., the fraction of classifiers that are considered when minimizing empirical error. Obviously, a large value of ρ means that more emphasis on the empirical error, while a small ρ pays more attention on the diversity.

5 Empirical Studies

In this section, we perform experiments to evaluate the proposed DREP method, also to validate the theoretical results.

5.1 Settings

In experiments, we use twenty binary classification data sets from the UCI repository [10], amongst which four data sets are generated from multi-class data sets, that is, *letter** classifies 'u' against 'v' on *letter*; *optdigits* classifies '01234' against '56789' on *optdigits*; *satimage** classifies labels '1' and '2' against those with '5' and '7' on *satimage*; and *vehicle** classifies 'bus' and 'opel' against 'van' and 'saab' on *vehicle*. Since these data sets are widely used benchmarks, we omit their summary information for clarity here.

The DREP method and several comparative methods are evaluated in a series of experiments. Specifically, each experiment is performed on one data set, and mainly involves the following steps:

1. Randomly split the data set into three parts: 1/3 as training set, 1/3 as validation set and the rest as test set;
2. Using Bagging [3] to build an ensemble of 100 CART decision trees [4] on the training set;
3. Prune the obtained ensemble by using ensemble pruning methods, whose parameters are determined on the validation set;
4. Evaluate the performance of pruned ensemble on the test set, also record the size of the pruned ensemble.

On each data set, each experiment is run for thirty times. At each time, the sizes of pruned ensemble and its error rates on test set are recorded, and finally the averaged results with standard deviation over multiple runs are reported.

In experiments, the comparative methods include two benchmark methods:

- Bagging [3]: it is the full ensemble of all the 100 CART trees;
- Best Individual (BI): it selects the individual classifier which has the best performance on the validation set.

Moreover, the following greedy forward ensemble pruning methods are implemented and compared:

- Reduce-Error (RE) [19, 6]: it starts with the classifier with lowest error, and then greedily selects the classifier that reduces error most;
- Kappa [19, 1]: it starts with the pair of classifiers with lowest κ-statistics, and then iteratively adds the classifier with lowest κ-statistics with respect to current sub-ensemble;
- Complementarity (CP) [21, 20]: this method starts with the classifier with lowest error, it incorporates at each iteration the one which is most complementary to the sub-ensemble;

Table 1. Error rates (mean±std.) achieved by comparative methods. On each data set an entry is marked with bullet '•' (or circle '○') if it is significantly better (or worse) than unpruned Bagging based on paired t-test at the significance level 0.1; the win/tie/loss counts are summarized in the last row.

Data set	Bagging	BI	RE	Kappa	CP	MD	DREP
australian	.134±.019	.148±.023○	.133±.015	.138±.017	.130±.014	.133±.016	.129±.016
breast-cancer	.278±.043	.293±.051	.275±.035	.280±.033	.288±.031	.311±.052○	.265±.024•
breast-w	.040±.010	.050±.012○	.034±.009•	.041±.011	.036±.009•	.035±.008	.034±.008•
diabetes	.239±.023	.256±.023○	.236±.022	.249±.020○	.240±.020	.243±.020	.234±.017
germen	.247±.016	.292±.021○	.248±.021	.254±.022	.250±.017	.247±.019	.248±.015
haberman	.261±.026	.270±.030○	.257±.025	.258±.034	.267±.029	.283±.037○	.252±.021
heart-statlog	.204±.039	.226±.041	.194±.034	.203±.035	.188±.034•	.195±.033	.183±.027•
hepatitis	.165±.030	.206±.049○	.164±.037	.183±.029○	.162±.031	.170±.036	.159±.027
ionosphere	.088±.024	.106±.034○	.069±.018•	.089±.030	.070±.019•	.079±.024	.066±.017•
kr-vs-kp	.014±.005	.013±.005	.009±.002•	.017±.005	.012±.004•	.015±.004	.008±.002•
*letter**	.047±.009	.075±.013○	.039±.008•	.047±.008	.039±.008•	.044±.008	.035±.007•
liver-dis	.313±.030	.362±.041○	.312±.032	.327±.039	.313±.035	.323±.038	.311±.029
*optdigits**	.046±.005	.109±.008○	.041±.004•	.045±.005	.040±.004•	.044±.005	.040±.003•
*satimage**	.032±.004	.051±.007○	.031±.004	.033±.005	.030±.004•	.032±.004	.029±.004•
sick	.016±.003	.017±.004	.015±.003	.017±.003	.015±.003	.016±.003	.014±.002•
sonar	.245±.050	.285±.036○	.235±.044	.245±.051	.216±.038•	.233±.044	.230±.030•
spambase	.071±.005	.094±.007○	.066±.005•	.070±.004	.066±.004•	.069±.005•	.066±.004•
tic-tac-toe	.060±.018	.101±.021○	.039±.008•	.082±.026○	.043±.010•	.078±.022○	.038±.007•
*vehicle**	.207±.020	.235±.029○	.207±.021	.214±.023	.204±.019	.215±.025	.203±.019
vote	.038±.011	.043±.014	.035±.011	.041±.016	.035±.013	.037±.011	.033±.007•
win/tie/loss	–	0/5/15	7/13/0	0/17/3	10/10/0	1/16/3	13/7/0

– Margin Distance (MD) [21, 20]: at each iteration, this method incorporates into the ensemble tho classifier that reduces the distance from the margin vector to the objective point in the first quadrant the most.

It is easy to find that RE and Kappa consider only empirical error and diversity respectively, while CP, MD and DREP take both of them into account. For each ensemble pruning method, we will stop and return the sub-ensemble if its error rate on validation set cannot be reduced. In the experiments, the κ-statistics, complementarity measure, margin distance are estimated on the validation set, and the parameter ρ of DREP is selected in $\{0.2, 0.25, \ldots, 0.5\}$ on validation set.

All the experiments are run on a PC with 2GB memory.

5.2 Results

The error rates achieved by comparative methods are shown in Table 1. On each data set, paired t-test at significance level 0.1 is performed to compare performance of BI and ensemble pruning methods with that of Bagging. In Table 1, an entry is marked with bullet '•' (or circle '○') if it is significantly better (or worse) than Bagging, and the win/tie/loss counts are summarized in the last row. From the results, it is shown that BI which selects the best performed individual loses at 15 out of 20 data sets against Bagging, this coincides with the fact that ensemble usually achieves better performance than a single classifier. Meanwhile,

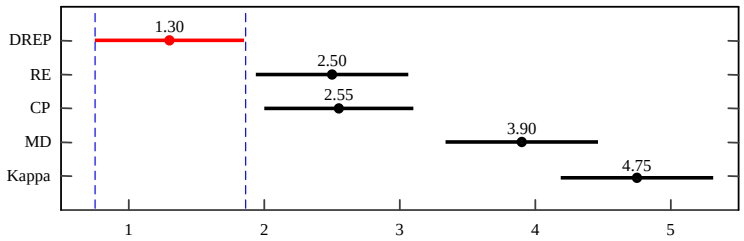

Fig. 1. The result of the Freidman test for comparing the performance of five ensemble pruning methods on 20 data sets. The dots indicate the average ranks, the bars indicate the critical difference with the Bonferroni-Dunn test at significance level 0.1, and compared methods having non-overlapped bars are significantly different.

the performance of ensemble pruning methods is much better. Specifically, RE, CP and DREP respectively achieve 7, 10 and 13 wins but no losses compared with Bagging, while Kappa and MD respectively make 17 and 16 ties and only 3 losses. At the same time, from Table 2 it can be seen that the ensemble sizes are reduced from 100 to about 20. Hence, the purpose of ensemble pruning (that is, reduce ensemble size whilst keeping or improving performance) is reached; also amongst comparative ensemble pruning methods, it appears that DREP method performs quite well (it achieves the best win/tie/loss counts and the smallest average ensemble size).

To better compare the performance of greedy ensemble pruning methods, we perform *Freidman test* [8], which is a non-parametric statistical significance test for comparing multiple classifiers on multiple data sets. Roughly speaking, the Freidman test is based on the ranks of compared methods on multiple data sets, and it is performed in conjunction with the *Bonferroni-Dunn test* at certain significance level. Here, we employ it to compare the five greedy ensemble pruning methods used in our experiments, and the result is shown in Fig. 1. Amongst the five pruning methods, the DREP method gets the highest average rank (1.30), followed by RE (2.50) and CP (2.55), while the average ranks of MD and CP are 3.90 and 4.75 respectively. Since the critical difference with the two-tailed Bonferroni-Dunn test for 5 classifiers on 20 data sets is $2.241\sqrt{(5 \cdot 6)/(6 \cdot 20)} \approx 1.121$, we can find that the performance of DREP is significantly better than other methods (in Fig. 1 the bar of DREP is not overlapping with either of other methods). It is easy to understand that the performance of DREP is better than that of RE and Kappa, because RE and Kappa only consider the empirical risk and the diversity respectively while DREP take both of them into account. Also, RE performs significantly better than Kappa, which may imply that empirical error play a more important role in the tradeoff. This coincides with our theoretical results, because diversity plays a role of regularization which is used to prevent overfitting, and only considering regularization usually does not help to improve the generalization performance. Furthermore, it can be seen that DREP performs better than CP and MD, this can be explained that DREP explicitly tradeoffs empirical error and diversity

Table 2. Ensemble sizes (mean±std.) of the pruned ensemble. On each data set, the entry achieving the smallest ensemble size is bolded, and the averaged sizes over all data sets are given in the last row.

Data set	RE	Kappa	CP	MD	DREP
australian	**15.4±3.5**	18.7±5.5	18.0±4.1	19.9±6.4	18.3±4.2
breast-cancer	18.4±3.8	22.0±9.2	18.1±5.1	24.1±10.0	**18.1±4.5**
breast-w	17.5±4.3	**15.5±3.9**	20.1±6.3	23.1±8.5	17.1±4.3
diabetes	23.5±5.8	26.7±11.4	21.9±6.4	29.0±10.1	**17.6±4.7**
germen	21.5±5.9	25.9±8.6	20.5±5.9	28.5±10.3	**17.1±3.9**
haberman	16.7±4.8	**15.7±5.0**	21.3±6.7	22.8±8.5	18.0±4.4
heart-statlog	18.9±4.1	22.9±8.1	21.9±5.8	23.5±7.8	**17.2±4.4**
hepatitis	14.4±2.6	**12.7±3.3**	17.7±5.3	18.9±5.9	17.7±4.6
ionosphere	**15.6±2.6**	21.8±9.0	19.5±5.8	23.8±7.7	17.5±4.6
kr-vs-kp	**16.1±3.7**	22.2±10.0	23.1±5.8	21.5±6.6	17.7±4.1
*letter**	22.9±4.6	25.0±7.6	**22.0±5.2**	27.8±9.2	24.7±4.5
liver-dis	22.7±5.7	23.6±10.9	21.5±5.6	25.5±9.3	**18.5±4.6**
*optdigits**	31.9±6.6	37.4±10.8	28.1±5.0	37.7±10.7	**25.1±4.8**
*satimage**	25.1±5.5	32.5±9.6	23.3±5.8	30.9±9.0	**24.8±4.7**
sick	**15.8±3.2**	22.7±10.9	20.9±4.7	22.8±9.0	17.7±4.4
sonar	20.1±5.0	22.1±9.8	21.1±5.9	24.9±9.5	**18.7±4.9**
spambase	25.3±6.8	27.5±8.6	24.3±7.0	28.4±8.3	**18.1±4.7**
tic-tac-toe	24.2±4.4	36.4±17.1	24.1±5.9	38.3±16.4	**18.9±4.3**
*vehicle**	**22.0±5.3**	25.7±8.5	22.3±7.1	25.2±9.4	24.9±4.9
vote	**12.8±1.6**	15.9±5.7	16.6±5.3	18.8±6.0	18.3±4.3
average	20.0	23.6	21.3	25.8	**19.3**

regularization, while CP and MD implicitly consider the tradeoff at a fixed level and can be easily affected by noises.

Table 2 presents the sizes of pruned ensembles, which shows that all the greedy pruning methods heavily reduce the ensemble sizes. Moreover, it can be seen that DREP achieves the smallest sizes on 10 data sets, also the smallest average ensemble size.

Furthermore, Fig. 2 plots the test error curves of Bagging and compared ensemble pruning methods on *heart-statlog* and *letter**. In detail, for Bagging the individual classifiers are aggregated in random order, and for ensemble pruning methods the greedy selection process will not be stopped until all the individuals are included, that is, the individual classifiers are aggregated in an order specified by the pruning methods. At each ensemble size the error is estimated on the test data, and the final results are obtained by averaging results of thirty runs, and they are plotted against ensemble sizes in Fig. 2. It can be seen that as ensemble size increases, the test error of Bagging decreases and converges, but the test errors of greedy ensemble pruning methods decrease much faster and are lower than Bagging, which indicates that better performance can be achieved at smaller ensemble sizes by using greedy ensemble pruning methods. By comparing the curves of DREP and other pruning methods, we can find that the test error of DREP decrease faster than other methods, even faster than RE which selects individual classifiers based on empirical error on the validation data set. This is not hard to understand because RE may overfit the validation

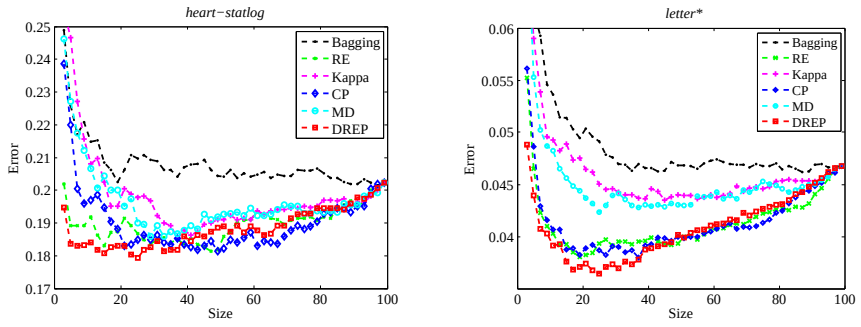

Fig. 2. Averaged test errors curves of Bagging and compared ensemble pruning methods on (a) *heart-statlog* and (b) *letter**, where horizontal axis and vertical axis correspond to ensemble size and test error respectively. For ensemble pruning methods, we do not stop the greedy selection process until all the individual classifiers are included.

data, while the diversity regularization used by DREP tends to help it achieve better performance.

In summary, we can see that with the help of diversity regularization, DREP is able to achieve significantly better generalization performance with smaller ensemble size than the compared methods.

6 Conclusion and Future Work

In ensemble learning, understanding diversity is one of the most important fundamental issues. This work focuses on the most popular setting of ensemble pruning, where the individual classifiers are combined by voting and the diversity is measured in the pairwise manner. In the PAC-learning framework, it presents a theoretical analysis on the role of diversity in voting, which is, to our best knowledge, the first PAC-style analysis on the effect of diversity in voting. It discloses that by enforcing large diversity, the hypothesis space complexity of voting can be reduced, and then better generalization performance can be expected. In the view of statistical learning, this implies that encouraging diversity can be regarded to apply regularization on ensemble methods. This may introduce a novel perspective of diversity in ensemble learning. Guided by this result, a greedy ensemble pruning method called DREP is proposed to explicitly exploit diversity regularization. Experimental results show that with the help of diversity regularization, DREP is able to achieve significantly better generalization performance with smaller ensemble size than the compared methods.

The current work applies diversity regularization on greedy ensemble pruning, it will be an interesting future work to develop ensemble learning methods which explicitly exploits diversity regularization. Recently it has been found that the ensemble diversity exists at multiple orders of correlation [5, 34], thus it is also of great interest to study whether the theoretical results on diversity still hold in that case.

Acknowledgements

The authors want to thank anonymous reviewers for helpful comments. This research was supported by the NSFC (61021062, 60903103) and the 973 Program (2010CB327903).

References

1. R. Banfield, L. Hall, K. Bowyer, and W. Kegelmeyer. Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1):49 – 62, 2005.
2. P. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Neural Networks*, 44(2):525–536, 1998.
3. L. Breiman. Bagging predictors. *Machine Learning*, 24(3):123–140, 1996.
4. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.
5. G. Brown. An information theoretic perspective on multiple classifier systems. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 344–353, Reykjavik, Iceland, 2009.
6. 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*, pages 18–25, 2004.
7. H. Chen, P. Tiňo, and X. Yao. Predictive ensemble pruning by expectation propagation. *IEEE Transactions on Knowledge and Data Engineering*, 21(7):999–1013, 2009.
8. J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, Dec. 2006.
9. T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.
10. A. Frank and A. Asuncion. UCI machine learning repository, 2010.
11. G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):942–956, 2005.
12. G. Giacinto, F. Roli, and G. Fumera. Design of effective multiple classifier systems by clustering of classifiers. In *Proceedings of the 15th International Conference on Pattern Recognition*, pages 160–163, Barcelona,Spain, 2000.
13. D. Hernández-Lobato, G. Martínez-Muñoz, and A. Suárez. Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles. *Neurocomputing*, 74(12-13):2250–2264, 2011.
14. A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*, pages 231–238, Denver, CO, 1994.
15. L. Kuncheva and C. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.
16. L. Kuncheva, C. Whitaker, C. Shipp, and R. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6(1):22–31, 2003.

17. A. Lazarevic and Z. Obradovic. Effective pruning of neural network classifier ensembles. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, pages 796–801, Washington, DC, 2001.
18. N. Li and Z.-H. Zhou. Selective ensemble under regularization framework. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 293–303, Reykjavik, Iceland, 2009.
19. D. Margineantu and T. Dietterich. Pruning adaptive boosting. In *Proceedings of the 14th International Conference on Machine Learning*, pages 211–218, Nashville, TN, 1997.
20. G. Martínez-Muñoz, D. Hernández-Lobato, and A. Suárez. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):245 –259, 2009.
21. G. Martínez-Muñoz and A. Suárez. Aggregation ordering in bagging. In *Proceeding of the IASTED International Conference on Artificial Intelligence and Applications*, pages 258–263, Innsbruck, Austria, 2004.
22. G. Martínez-Muñoz and A. Suárez. Pruning in ordered bagging ensembles. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 609–616, Pittsburgh, PA, 2006.
23. I. Partalas, G. Tsoumakas, and I. Vlahavas. Focused ensemble selection: A diversity-based method for greedy ensemble selection. In *Proceedings of 18th European Conference on Artificial Intelligence*, pages 117–121, Patras, Greece, 2008.
24. I. Partalas, G. Tsoumakas, and I. Vlahavas. A study on greedy algorithms for ensemble pruning. Technical Report TR-LPIS-360-12, Department of Informatics, Aristotle University of Thessaloniki, Greece, 2012.
25. R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.
26. C. Tamon and J. Xiang. On the boosting pruning problem. In *Proceedings of the 11th European Conference on Machine Learning*, pages 404–412, Barcelona, Spain, 2000.
27. E. K. Tang, P. Suganthan, and X. Yao. An analysis of diversity measures. *Machine Learning*, 65(1):247–271, 2006.
28. G. Tsoumakas, I. Partalas, and I. P. Vlahavas. An ensemble pruning primer. In O. Okun and G. Valentini, editors, *Applications of Supervised and Unsupervised Ensemble Methods*, pages 1–13. Springer, 2009.
29. L. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.
30. Y. Yu, Y.-F. Li, and Z.-H. Zhou. Diversity regularized machine. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1603–1608, Barcelona, Spain, 2011.
31. T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(March):527–550, 2002.
32. Y. Zhang, S. Burer, and W. Street. Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research*, 7:1315–1338, 2006.
33. Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, Boca Raton, FL, 2012.
34. Z.-H. Zhou and N. Li. Multi-information ensemble diversity. In *Proceedings of the 9th International Workshop on Multiple Classifier Systems*, pages 134–144, Cairo, Egypt, 2010.
35. Z.-H. Zhou, J. Wu, and W. Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002.