When does Diversity Help Generalization in Classification Ensembles?

Yijun Bian, and Huanhuan Chen, *Senior Member, IEEE*

***Abstract*—Ensembles, as a widely used and effective technique in the machine learning community, succeed within a key element—"diversity." The relationship between diversity and generalization, unfortunately, is not entirely understood and remains an open research issue. To reveal the effect of diversity on the generalization of classification ensembles, we investigate three issues on diversity, i.e., the measurement of diversity, the relationship between the proposed diversity and generalization error, and the utilization of this relationship for ensemble pruning. In the diversity measurement, we measure diversity by error decomposition inspired by regression ensembles, which decomposes the error of classification ensembles into accuracy and diversity. Then we formulate the relationship between the measured diversity and ensemble performance through the theorem of margin and generalization, and observe that the generalization error is reduced effectively only when the measured diversity is increased in a few specific ranges, while in other ranges larger diversity is less beneficial to increase generalization of an ensemble. Besides, we propose a pruning method based on diversity management to utilize this relationship, which could increase diversity appropriately and shrink the size of the ensemble with non-decreasing performance. The experiments validate the effectiveness of this proposed relationship between the proposed diversity and the ensemble generalization error.**

***Index Terms*—ensemble learning, diversity, ensemble pruning, error decomposition.**

I. Introduction

Ensemble learning has attracted plenty of research attention in the machine learning community thanks to its remarkable potential [1], and it has been widely used in many real-world applications such as object detection, object recognition, and object tracking [2]–[5]. Rather than relying on one single model, an ensemble is a set of learned models that make decisions collectively. It is widely accepted that an ensemble usually generalizes better than one single model [6]–[9]. Dietterich [10] stated that an ensemble of classifiers succeeded in better accuracy if and only if its individual members are accurate and diverse. One classifier is accurate if its error rate is better than random guessing on new instances. Two classifiers are diverse if they make different errors in new instances. The diversity of an ensemble usually decreases with the increasing of accuracies of its members [11]. Thus, how to handle the trade-off between accuracy and diversity appropriately is a crucial issue in ensemble learning.

Unfortunately, there is still no consensus in the community on the definition or measurement for diversity, unlike the precise accuracy. Existing ensemble methods create different classifiers implicitly or heuristically by manipulating input data, input features, or output targets. Similarly, diversity could be measured in either the input space (i.e., instance features) or the output space (i.e., classification results) [9]. Existing diversity measures are various, yet none of them show the superiority over each other [12]. They are generally divided into pairwise and non-pairwise diversity measures [13]. Lots of previous work made an effort to seek the role of diversity in ensemble learning [14]–[17]. In regression ensembles, the diversity is defined based on the error decomposition [18], in which the error of regression ensembles is split into the accuracy term and diversity term. The classic Ambiguity Decomposition [18] and Bias-Variance-Covariance decomposition [19] are two commonly used error decomposition schemes. However, both of them are only suitable for regression tasks with the square loss.

Inspired by error decomposition of regression ensembles, we propose a measure of diversity using error decomposition for classification ensembles with the 0/1 error function firstly, where the error of classification ensembles is split into two terms: accuracy and diversity. Like regression ensembles, the diversity term measures the difference among members of classification ensembles. Secondly, we propose the relationship between the proposed diversity and generalization using the proposed diversity measure based on [20], [21]. By taking bagging and AdaBoost as examples of ensemble methods, experiments are conducted to validate the proposed relationship by varying the diversity in bagging and AdaBoost. Thirdly, we propose an ensemble pruning method named as Ensemble Pruning based on Diversity (*EPBD*), which utilizes the proposed relationship between the proposed diversity and generalization to improve the ensemble's generalization. Moreover, empirical results are presented to validate the effectiveness of *EPBD*. The contributions in this paper are three-fold:

- A diversity measure in classification ensembles is proposed to quantify the difference among ensemble members based on the error decomposition in classification ensembles, with the benefits to conduct theoretical analyses about its properties.
- The theoretical analyses on the relationship between the proposed diversity measure and the generalization of the ensemble is investigated, which demonstrates that diversity has different impacts to the ensemble generalization in different regions.
- An ensemble pruning method, i.e., *EPBD*, is proposed to select a subset of the original ensemble by utilizing the

Y. Bian and H. Chen are with the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei 230027, China (e-mails: yjbian@mail.ustc.edu.cn; hchen@ustc.edu.cn).

theoretical relationship between the proposed diversity and the ensemble generalization error.

The rest of this paper is organized as follows. The related work is summarized in Section II. And then, the core investigations are presented in Section III, answering three issues about diversity: (1) the measurement of diversity and its corresponding properties in Section III-A and III-B; (2) the relationship between the proposed diversity and generalization error of ensembles in Section III-C; (3) the utilization of this relationship about diversity in Section III-D. Finally, the empirical results are presented in Section IV, followed by the conclusion in Section V.

II. Related Work

Diversity is considered intuitively as the difference among individual members in an ensemble, with several alternative names as dependence, orthogonality or complementarity [1], [13]. In this section, we introduce existing methods to generate diverse individual classifiers. Then we describe existing methods to measure diversity in ensemble learning. Finally, we summarize some research about the crucial role diversity plays in ensemble learning.

A. Diversity Generation

Most ensemble methods attempt to generate diverse classifiers implicitly or heuristically by manipulating the input data or features, while few of them manipulate output targets [10], [16]. For example, bagging [22] and Boosting (including many variants) [23]–[25] manipulate input data to promote diversity by choosing different subsets of the training data during training; Random forest [26] manipulates subsets of input data or features to create diversity and gives competitive results; Rotation forest [27] applies principal component analysis (PCA) on each subset as an improved method. Moreover, neural network (NN) ensembles also create diversity using different initial weights, different architectures of the networks, and different learning algorithms.

B. Diversity Measures

Diversity can be measured in either the input space or the output space [9]. Most work focuses on output diversity [12], [15], [28]–[30] while only few focuses on input diversity [31]. Existing diversity measures are generally divided into pairwise and non-pairwise diversity measures [13]. Based on the coincident errors between pairs of individual classifiers, pairwise diversity represents the behavior if both of them predict an instance identically or disagree with each other. In this case, the overall diversity is the average of all possible pairs [9]. Pairwise diversity includes Q-statistic [32], κ-statistic [33], disagreement measure [34], [35], correlation coefficient [36], and double-fault [37]. In contrast, non-pairwise diversity directly measures a set of classifiers using the variance, entropy, or the proportion of individual classifiers that fail on randomly chosen instances [9]. It includes interrater agreement [38], Kohavi-Wolpert variance [39], the entropy of the votes [40], [41], the difficulty index [13], [42], the generalized diversity [43], and the coincident failure diversity [43]. Apart from those, two other measures exist and do not fall into the categories above. One is the correlation penalty function, proposed by Liu and Yao [44] in their negative correlation learning (NCL) [45]. It measures the diversity of each member against the entire ensemble. The other is ambiguity, which measures the average offset of each member against the entire ensemble output [46].

C. The Role that Diversity Plays

Although the crucial role of diversity has been widely accepted, few researchers can tell how diversity works exactly in ensemble methods. In the last decade or so, Brown [14] claimed that from an information-theoretic perspective, diversity within an ensemble did exist on numerous levels of interaction between the classifiers. This work inspired Zhou and Li [15] to propose that the mutual information should be maximized to minimize the prediction error of an ensemble from the view of multi-information. Subsequently, Yu *et al.* [16] claimed that the diversity among individual learners in a pairwise manner, used in their diversity regularized machine (DRM), could reduce the hypothesis space complexity, which implied that controlling diversity played the role of regularization in ensemble methods. Recently, Jiang *et al.* [17] extended the classic Ambiguity Decomposition from regression problems to classification problems and proved several fundamental properties of the Ambiguity term.

III. Methodology

In this section, we formally study the measure of diversity using error decomposition of classification ensembles and derive proper theoretical analyses from quantifying and characterizing the relationship between the proposed diversity and generalization performance of the entire ensemble. Based on the results of the analyses, an ensemble pruning method is proposed to construct a pruned sub-ensemble effectively.

We first introduce the necessary notations used in this paper. Let vectors be denoted by bold lowercase letters (e.g., $\mathbf{x} = (x_1, ..., x_k)$) as an instance with k-tuples of real numbers describing features, and scalars denoted by italic lowercase letters (e.g., y) as labels. Random variables, (vector) spaces, and subsets of these vector spaces are denoted by sans serif uppercase letters (e.g., X), calligraphic uppercase letters (e.g., $\mathcal{X}$), and roman uppercase letters (e.g., W), respectively. Sets of vectors are denoted by roman uppercase letters (e.g., S) consisting of some instances with their corresponding labels; individual classifiers/learners are denoted by functions (e.g., $f(\cdot)$) as the output results of one classifier. The symbols $\mathbf{P}, \mathbf{E}, \mathbb{I}$, and $\mathbb{R}$ denote the probability measure, the expectation of a random variable, the indicator function, and the real space, respectively. The notation $(\mathbf{x}, y) \in S$ is formally defined by $(\mathbf{x}, y) \in S \Leftrightarrow \exists i \in \{1, ..., |S|\} : (\mathbf{x}_i, y_i) = (\mathbf{x}, y)$ for clarity.

A. Measuring Diversity in Classification Ensembles using Error Decomposition

The analysis is built in the PAC (probably approximately correct) framework [47] that one learning task $\mathcal{D}$ corresponds to a probability distribution over the input-output space $\mathcal{X}\times\mathcal{Y}$. An instance from $\mathcal{D}$ is represented as $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^k$ is a vector representing features and y is a scalar as a label. Suppose this classification task is to use an ensemble that comprises several individual classifiers to approximate a function $f_{true} : \mathcal{X} \mapsto \mathcal{Y}$, and $F = \{f_1, ..., f_{|F|}\}$ denotes the set of these individual classifiers. The predictions of the individual classifiers are combined in the spirit of weighted averaging and plurality voting[1], that is, the output of an ensemble is defined as:

$$f_{ens}(\mathbf{x}) = \operatorname{sgn}\left(\sum_{f\in F} c\cdot f(\mathbf{x})\right), \tag{1}$$

where c is the coefficient corresponding to the individual classifier f. And $\sum_{f\in F} c\cdot f(\mathbf{x}) = 0$ indicates that there is a tie in the combination.

In this work, we focus on binary classification problems that means $\mathcal{Y} = \{-1, +1\}$, but these theoretical results could be expanded to multi-classification problems. Assume that there is a training set S with several instances $(\mathbf{x}, y)$. For one single arbitrary instance $\mathbf{x} \in S$, y denotes the target output of this instance, and $f(\mathbf{x})$ is the actual output of the individual classifier $f \in F$. Notice that y and $f(\mathbf{x})$ satisfy that $y, f(\mathbf{x}) \in \mathcal{Y}$. If the actual output of the individual classifier is correct according to the target output, obviously $f(\mathbf{x})\cdot y = +1$, otherwise $f(\mathbf{x}) \cdot y = -1$, where this term is defined as the *margin* of this classifier on the instance,

$$\operatorname{margin}(f, \mathbf{x}) = f(\mathbf{x}) \cdot y\,. \tag{2}$$

As the original error decomposition in regression [18], [48], [49] uses the square loss as the loss function that is not suitable for classification, we employ the 0/1 error function to adopt the idea of error decomposition for regression ensembles. For classification problems, the error function of a classifier f at a single arbitrary instance $\mathbf{x}$ is defined as

$$\operatorname{Err}(f(\mathbf{x})\cdot y) = \begin{cases} 1\,, & \text{if } f(\mathbf{x})y = -1\,;\\ 0.5\,, & \text{if } f(\mathbf{x})y = 0\,;\\ 0\,, & \text{if } f(\mathbf{x})y = 1\,, \end{cases} \tag{3}$$

which is also the discretization of hinge loss function and holds that $\operatorname{Err}(f(\mathbf{x})y) = -\frac{1}{2}(f(\mathbf{x})y - 1)$. Therefore, inspired by the work of [17], [48], we present the following error decomposition for classification ensembles.

Theorem 1 (Error decomposition for classification ensembles)**.** *Assume that we are dealing with binary classification problems. Individual classifiers in an ensemble $F = \{f_1, ..., f_{|F|}\}$ have been trained and are combined by weighted plurality voting $f_{ens} = \operatorname{sgn}\left(\sum_{f\in F} c\cdot f(\mathbf{x})\right)$ with $\sum_{f\in F} c = 1$. Then for the 0/1 error function, the loss function of the ensemble can be decomposed into*

$$\begin{aligned}\operatorname{Err}(f_{ens}(\mathbf{x})y) = &\sum_{f\in F} c\cdot \operatorname{Err}(f(\mathbf{x})y)\\ &-\frac{1}{2}\Big(\operatorname{margin}(f_{ens}, \mathbf{x}) - \sum_{f\in F} c\cdot\operatorname{margin}(f, \mathbf{x})\Big)\,.\end{aligned} \tag{4}$$

Computing the expectation in the instance space yields decomposition of the generalization error as follows,

$$\bar{G} = \bar{A} - \bar{D}\,, \tag{5}$$

where

$$\bar{G} = \mathbf{E}_S(\operatorname{Err}(f_{ens}(\mathbf{x})y))\,, \tag{6a}$$

$$\bar{A} = \sum_{f\in F} c\cdot\mathbf{E}_S(\operatorname{Err}(f(\mathbf{x})y))\,, \tag{6b}$$

$$\bar{D} = \frac{1}{2}\mathbf{E}_S(\operatorname{margin}(f_{ens}, \mathbf{x})) - \frac{1}{2}\sum_{f\in F} c\cdot\mathbf{E}_S(\operatorname{margin}(f, \mathbf{x}))\,, \tag{6c}$$

with $\mathbf{E}_S(\cdot)$ representing the expectation in the instance space.

Proof. For one single arbitrary instance $\mathbf{x}$, the error of an ensemble is defined as

$$\operatorname{Err}(f_{ens}(\mathbf{x})\cdot y) = \operatorname{Err}\left(\operatorname{sgn}\Big(\sum_{f\in F} c\cdot f(\mathbf{x})\Big)\cdot y\right). \tag{7}$$

According to the error decomposition of regression ensembles, the difference between ensemble error and the average error of individual classifiers is

$$\begin{aligned}&\operatorname{Err}(f_{ens}(\mathbf{x})\cdot y) - \sum_{f\in F} c\cdot\operatorname{Err}(f(\mathbf{x})\cdot y)\\ =&\sum_{f\in F} c\cdot\Big(\operatorname{Err}(f_{ens}(\mathbf{x})\cdot y) - \operatorname{Err}\big(f(\mathbf{x})\cdot y\big)\Big)\\ =&-\frac{y}{2}\sum_{f\in F} c\cdot\big(f_{ens}(\mathbf{x}) - f(\mathbf{x})\big)\,,\end{aligned} \tag{8}$$

and the error decomposition could be rewritten as

$$\operatorname{Err}(f_{ens}(\mathbf{x})y) = \sum_{f\in F} c\cdot\operatorname{Err}(f(\mathbf{x})y) - \frac{1}{2}\sum_{f\in F} c\big(f_{ens}(\mathbf{x}) - f(\mathbf{x})\big)y\,. \tag{9}$$

The first term, $\sum_{f\in F} c\cdot\operatorname{Err}(f(\mathbf{x})y)$, is the weighted average loss of the individuals. The second term is defined as the *diversity* term which measures the difference between $f_{ens}(\mathbf{x})$ and individual classifiers, and it could be rewritten as

$$\begin{aligned}&\frac{1}{2}\sum_{f\in F} c\big(f_{ens}(\mathbf{x}) - f(\mathbf{x})\big)y\\ =&\frac{1}{2}\sum_{f\in F} c\Big(\operatorname{margin}(f_{ens}, \mathbf{x}) - \operatorname{margin}(f, \mathbf{x})\Big)\,.\end{aligned} \tag{10}$$

Note that this term is corresponding to one specific instance $\mathbf{x}$, defined as

$$\operatorname{div}(f_{ens}, \mathbf{x}) = \frac{1}{2}\operatorname{margin}(f_{ens}, \mathbf{x}) - \frac{1}{2}\sum_{f\in F} c\cdot\operatorname{margin}(f, \mathbf{x})\,. \tag{11}$$

And then we obtain the form of the error decomposition for one single instance $\mathbf{x}$ as in Eq. (4). Finally, computing the expectation over this instance space will yield the form of the error decomposition for this overall instance set as in Eq. (5). □

[1]In plurality voting, each individual classifier votes for a class and the class label receiving the most number of votes is regarded as the output of the ensemble. If there is a tie, the Err function, Eq. (7) gives zero for binary classification.

Up to now, the diversity measure in Eq. (11) is proposed based on the error decomposition for classification ensembles, which would serve as the independent variable in the relationship between the proposed diversity and ensemble performance in Section III-C. Before that, the properties of the proposed diversity will be investigated in the next subsection.

B. Properties of $\bar{G}, \bar{A}, \bar{D}$

In this section, we analyze $\bar{G}, \bar{A}, \bar{D}$ in Eq. (5) of Theorem 1 to further explore the properties of these three terms.

Corollary 2 ($\bar{G}$, $\bar{A}$, $\bar{D}$ only depend on $\mathrm{div}(f_{ens}, \mathbf{x})$)**.** *Assume that we are dealing with binary classification problems. Individual classifiers in an ensemble $F = \{f_1, ..., f_{|F|}\}$ have been trained and are combined by weighted plurality voting $f_{ens} = \mathrm{sgn}(\sum_{f\in F} c \cdot f(\mathbf{x}))$ with $\sum_{f\in F} c = 1$. Then the generalization error $\bar{G}$, accuracy $\bar{A}$ and diversity $\bar{D}$ of this ensemble are all dependent on $\mathrm{div}(f_{ens}, \mathbf{x})$, that is,*

$$\bar{G} =\frac{1}{2}(1-\mathbf{E}_S(\mathrm{sgn}(\lambda-2\,\mathrm{div}(f_{ens},\mathbf{x})))) \,, \tag{12a}$$

$$\bar{A} =\frac{1}{2}(1-\mathbf{E}_S(\lambda-2\,\mathrm{div}(f_{ens},\mathbf{x}))) \,, \tag{12b}$$

$$\bar{D} =\mathbf{E}_S(\mathrm{div}(f_{ens},\mathbf{x})) \,, \tag{12c}$$

where

$$\lambda = \begin{cases} 1\,, & \text{if } \mathrm{div}(f_{ens},\mathbf{x}) \in (0,\frac{1}{2})\,; \\ 0\,, & \text{if } \mathrm{div}(f_{ens},\mathbf{x}) = 0\,; \\ -1\,, & \text{if } \mathrm{div}(f_{ens},\mathbf{x}) \in (-\frac{1}{2},0)\,. \end{cases} \tag{13}$$

Proof. The error of one individual classifier is

$$\mathrm{Err}(f(\mathbf{x})y) = -\frac{1}{2}(f(\mathbf{x})y-1) = -\frac{1}{2}(\mathrm{margin}(f,\mathbf{x})-1) \,, \tag{14}$$

as described previously, therefore, according to Theorem 1, we could obtain that

$$\bar{G} =\frac{1}{2}(1-\mathbf{E}_S(\mathrm{margin}(f_{ens},\mathbf{x}))) \,, \tag{15a}$$

$$\bar{A} =\frac{1}{2}\Big(1-\mathbf{E}_S\Big(\sum_{f\in F} c\cdot \mathrm{margin}(f,\mathbf{x})\Big)\Big) \,, \tag{15b}$$

$$\bar{D} =\frac{1}{2}\sum_{f\in F} c\cdot \mathbf{E}_S(\mathrm{margin}(f_{ens},\mathbf{x})-\mathrm{margin}(f,\mathbf{x})) \,. \tag{15c}$$

On the other side, according to Eq. (11), we could obtain that

$$\begin{aligned} &\mathrm{div}(f_{ens},\mathbf{x}) \\ =&\frac{1}{2}\,\mathrm{sgn}\left(\sum_{f\in F} c\cdot f(\mathbf{x})\right)y - \frac{1}{2}\sum_{f\in F} c\cdot \mathrm{margin}(f,\mathbf{x}) \\ =&\frac{1}{2}\,\mathrm{sgn}\left(\sum_{f\in F} c\cdot \mathrm{margin}(f,\mathbf{x})\right) - \frac{1}{2}\sum_{f\in F} c\cdot \mathrm{margin}(f,\mathbf{x}) \,, \end{aligned} \tag{16}$$

then define that

$$\overline{\mathrm{margin}}(f_{ens},\mathbf{x}) \hat{=} \sum_{f\in F} c\cdot \mathrm{margin}(f,\mathbf{x}) \,, \tag{17}$$

and consequently obtain that for clarity,

$$\overline{\mathrm{margin}}(f_{ens},\mathbf{x}) = \lambda - 2\,\mathrm{div}(f_{ens},\mathbf{x}) \,, \tag{18}$$

where λ is defined in Eq. (13). Finally, we could obtain that Eq. (12a)–(12c) as described in Corollary 2. □

C. Relationship Between Diversity and Ensemble Performance

The relationship between diversity and generalization in the literature remains an open question. Some researchers, such as Kuncheva and Whitaker [13], hold doubts about the usefulness of diversity measures in building classification ensembles in real-life pattern recognition problems. In contrast, some researchers hold the view as well that diversity among the members of a team of classifiers is a crucial issue in classifiers' combination [29]. They are two contradicting views, so, weirdly, both of them have some supporting experimental results in some cases. For instances, some empirical results [13] showed less correlation between diversity and generalization error by varying the diversity in the ensemble. Meanwhile, some experiments [16] showed precisely the opposite results. Hence, to clarify of this issue, whether or not diversity affects the performance of generalization, becomes particularly essential. Herbrich and Graepel [20], [21] presented that the relationship between generalization and margin did exist. Based on their work and the relationship between the proposed diversity and margin, we propose the relationship between the proposed diversity and generalization in the PAC framework, as shown in Figure 1.

Given the input/feature space $\mathcal{X}$, the output/label space $\mathcal{Y}$, and the training set S as described above, any instance $(\mathbf{x}, y) \sim (\mathsf{X},\mathsf{Y}) \in (\mathcal{X},\mathcal{Y})$ in S is drawn independent and identically distributed (i.i.d.) according to a certain unknown probability measure $\mathbf{P}_{\mathsf{Y}|\mathsf{X}}\mathbf{P}_{\mathsf{X}}$, and we only consider linear classifiers as classification ensembles just like $f_{ens}(\cdot)$ in Eq. (1),

$$\mathcal{F} = \{\mathbf{x} \mapsto \mathrm{sgn}(\langle \mathbf{c}, \phi(\mathbf{x})\rangle) \,|\, \mathbf{c}\in\mathcal{C}\} \,, \tag{19}$$

where $\phi(\mathbf{x}) = [f_1(\mathbf{x}), ..., f_{|F|}(\mathbf{x})]^\mathsf{T}$ and $\mathbf{c} = [c_1, ..., c_{|F|}]^\mathsf{T}$. Notice that $\sum_{f\in F} c = 1$ leads to a one-to-one correspondence of hypotheses $f_{ens}(\mathbf{x},\mathbf{c}) \in \mathcal{F}$ to their parameters $\mathbf{c}\in\mathcal{C}$. Assume that the existence of a "true" hypothesis $\mathbf{c}^* \in \mathcal{C}$ labeling the data, leading to a PAC-likelihood,

$$\mathbf{P}_{\mathsf{Y}|\mathsf{X}=\mathbf{x}}(y) = \mathbb{I}(y = \mathrm{sgn}(\langle \mathbf{c}^*, \phi(\mathbf{x})\rangle)) \,. \tag{20}$$

There exists a version space $V(S) \subseteq \mathcal{C}$ because of the existence of $\mathbf{c}^*$,

$$V(S) = \{\mathbf{c}\in\mathcal{C} \,|\, \forall(\mathbf{x},y)\in S\,, \mathrm{sgn}(\langle \mathbf{c},\phi(\mathbf{x})\rangle) = y\} \,. \tag{21}$$

The true risk $R(\mathbf{c})$ of consistent hypotheses $\mathbf{c}\in V(S)$ is

$$R(\mathbf{c}) = \mathbf{E}_{\mathsf{XY}}[\mathbb{I}(\mathrm{sgn}(\langle \mathbf{c},\phi(\mathsf{X})\rangle) \neq \mathsf{Y})] \,, \tag{22}$$

and a concept of the *margin* $\gamma(\mathbf{c})$ of an ensemble $\mathbf{c}$ is introduced [20], [21] as

$$\gamma(\mathbf{c}) = \min_{(\mathbf{x},y)\in S} y\langle \mathbf{c},\phi(\mathbf{x})\rangle \,, \tag{23}$$

since all ensemble classifiers $\mathbf{c}\in V(S)$ are indistinguishable in terms of error rate on the given training set S . Herbrich and Graepel [21] presented that the generalization error $R(\mathbf{c})$ of an ensemble was bounded from above in terms of the margin $\gamma(\mathbf{c})$, as shown in Lemma 3.

Lemma 3 (Relationship between margin and generalization [21])**.** *For any probability measures $\mathbf{P}_{\mathsf{X}}$ such as $\mathbf{P}_{\mathsf{X}}(\|\phi(\mathbf{x})\| \leqslant \delta) = 1$ and for any $\xi \in (0,1]$, with probability at least $(1-\xi)$ over the random draw of the training instances*

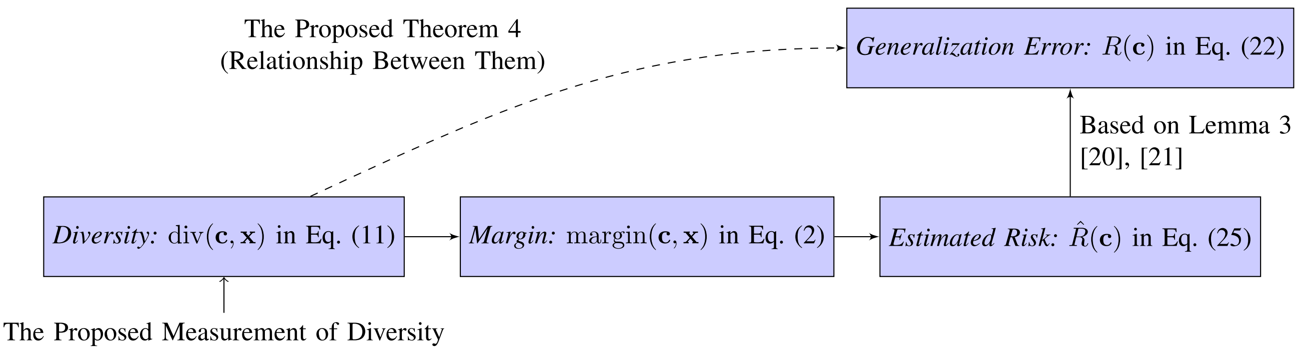


Fig. 1. Illustration for the motivation of the relationship between the proposed diversity and ensemble performance based on [20], [21]. Herbrich and Graepel [20], [21] proposed the relationship between margin and generalization (in Lemma 3). Besides, the proposed diversity in Eq. (11) is related to the margin. In this case, we could propose the relationship between the proposed diversity and ensemble performance (in Theorem 4). Note that the arrow from A to B means that B is related to A, where arrows in solid lines represent direct relations and that in the dotted line represents the indirect relation.

$S \in (\mathcal{X},\mathcal{Y})^{|S|}$ *, for any arbitrary consistent ensemble classifier* $\mathbf{c} \in V(S)$ *with a positive margin* $\gamma(\mathbf{c}) > \sqrt{32\delta^2/|S|}$*, the generalization error* $R(\mathbf{c})$ *could be bounded from above as following,*

$$R(\mathbf{c}) \leqslant \frac{2}{|S|}\left(\kappa(\mathbf{c})\log_2\left(\frac{8e|S|}{\kappa(\mathbf{c})}\right)\log_2(32|S|) + \log_2\left(\frac{2|S|}{\xi}\right)\right), \tag{24}$$

where $\kappa(\mathbf{c}) = \left\lceil \left(\frac{8\delta}{\gamma(\mathbf{c})}\right)^2 \right\rceil$.

Subsequently, inspired by the relationship between margin $\gamma(\mathbf{c})$ and generalization error R as in Lemma 3 [21], we present the following relationship between the proposed diversity and generalization. It is worth mentioning that the margin $\gamma(\mathbf{c})$ could be linked with the proposed diversity $\mathrm{div}(f_{ens},\mathbf{x})$ in Eq. (11), which exactly inspired us to pursue the relationship between the proposed diversity and the generalization risk. To simplify the analysis, we only consider the most important item in Eq. (24), i.e., $\kappa(\mathbf{c})\log_2\left(\frac{8e|S|}{\kappa(\mathbf{c})}\right)$, described as $\hat{R}(\mathbf{c})$.

Theorem 4 (Relationship between the proposed diversity and generalization)**.** *For any probability measures* $\mathbf{P}_{\mathsf{X}}$ *such as* $\mathbf{P}_{\mathsf{X}}(\|\phi(\mathbf{x})\| \leqslant \delta) = 1$ *and for any* $\xi \in (0,1]$ *, the random draw of the training instances* $S \in (\mathcal{X},\mathcal{Y})^{|S|}$ *is given with probability* ε *by noise disturbance. Then for any arbitrary consistent ensemble classifier* $\mathbf{c} \in V(S)$ *, with probability at least* $(1-\xi)$ *, the variation tendency of the upper bound of generalization error* $R(\mathbf{c})$ *is the same as that of* $\hat{R}(\mathbf{c})$ *, that is,*

$$\hat{R}(\mathbf{c}) = \left(\frac{8\delta}{\gamma(\mathbf{c})}\right)^2 \log_2\left(8e|S|\left(\frac{\gamma(\mathbf{c})}{8\delta}\right)^2\right), \tag{25}$$

$$\gamma(\mathbf{c}) = \min_{(\mathbf{x},y)\in S}(1-2\varepsilon)(\lambda - 2\,\mathrm{div}(f_{ens},\mathbf{x})), \tag{26}$$

where

$$\lambda = \begin{cases} 1\,, & \textit{if } \mathrm{div}(f_{ens},\mathbf{x}) \in (0,\frac{1}{2})\,; \\ 0\,, & \textit{if } \mathrm{div}(f_{ens},\mathbf{x}) = 0\,; \\ -1\,, & \textit{if } \mathrm{div}(f_{ens},\mathbf{x}) \in (-\frac{1}{2},0)\,. \end{cases} \tag{27}$$

Proof. Due to the existence of noise in real data, for any arbitrary instance $\mathbf{x}$ in S , let z be the observed label of this instance, and let y be the true unknown label corresponding. Thus, the error rate of the training set S is recorded as $\mathbf{P}(z \neq y \,|\, \mathbf{x} \in S) \hat{=} \varepsilon$, and the error of a classifier f is recorded as $\mathbf{P}(f(\mathbf{x}) \neq z \,|\, \mathbf{x} \in S) \hat{=} \theta$. Since the training set is affected by noise and other factors, we believe that $\theta \geqslant \varepsilon$. Then we use the observable z to estimate the value of y, i.e., $\hat{y} = (1-\varepsilon)z + \varepsilon(-z) = (1-2\varepsilon)z$. Thus, for a given training set $S = \{(\mathbf{x}_i, z_i) \,|\, i = 1,...,|S|\}$, individual classifiers are trained to constitute an ensemble, i.e., the set $F = \{f_1,...,f_{|F|}\}$. The margin of one individual classifier f_j on the instance $\mathbf{x}_i$ is

$$\mathrm{margin}(f_j,\mathbf{x}_i) = f_j(\mathbf{x}_i)z_i\,. \tag{28}$$

Describe $\overline{\mathrm{margin}}(f_{ens},\mathbf{x}_i)$ as in Eq. (17), and then according to Eq. (18) the margin of the ensemble is obtained

$$\gamma(\mathbf{c}) = \min_{1\leqslant i\leqslant |S|}(1-2\varepsilon)(\lambda - 2\,\mathrm{div}(f_{ens},\mathbf{x}_i)), \tag{29}$$

based on Eq. (23), which is mainly reflecting the difference among different members of the ensemble, i.e., diversity. Let $\mathbf{x}^*$ be the very instance where the ensemble reaches the minimum of the margin $\gamma(\mathbf{c})$, that is,

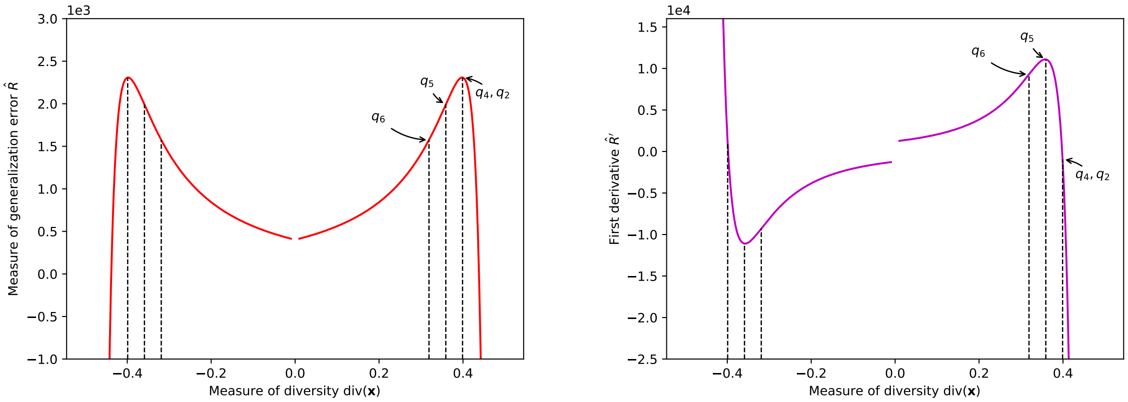
$$\gamma(\mathbf{c}) = (1-2\varepsilon)(\lambda - 2\,\mathrm{div}(f_{ens},\mathbf{x}^*)), \tag{30}$$

then $\kappa(\mathbf{c})$ is obtained based on Eq. (24). To simplify the analysis, we only consider the most important item in Eq. (24), i.e., $\kappa(\mathbf{c})\log_2\left(\frac{8e|S|}{\kappa(\mathbf{c})}\right)$, described as $\hat{R}(\mathbf{c})$, i.e.,

$$\begin{aligned} \hat{R}(\mathbf{c}) &\hat{=} \kappa(\mathbf{c})\log_2\left(\frac{8e|S|}{\kappa(\mathbf{c})}\right) \\ &\approx \left(\frac{8\delta}{(1-2\varepsilon)(\lambda - 2\,\mathrm{div}(f_{ens},\mathbf{x}^*))}\right)^2 \\ &\quad \cdot \log_2\left(8e|S|\left(\frac{(1-2\varepsilon)(\lambda - 2\,\mathrm{div}(f_{ens},\mathbf{x}^*))}{8\delta}\right)^2\right), \end{aligned} \tag{31}$$

with the same variation tendency as $R(\mathbf{c})$. □

According to Theorem 4, the generalization error of the ensemble could be quantified by the diversity. Eq. (31) reflects the relationship of individual classifiers $\mathrm{div}(f_{ens},\mathbf{x}^*) \in (-\frac{1}{2},\frac{1}{2})$ and the performance of generalization error of the en-



(a) (b)

Fig. 2. Illustration of the estimator of generalization error and its first derivative, impacted by our proposed measure of diversity. Notice that these endpoints (q_4, q_5 and q_6) in this figure refer to those points on the horizontal axis with the same abscissas correspondingly in fact. In this graph, parameters' values are specialized as follows: $\delta = 1$, $\varepsilon = 0.01$, $|S| = 200$. (a) Relationship between diversity and the ensemble generalization. (b) Relationship between diversity and the first derivative of generalization.

semble $R(\mathbf{c})$. Therefore, the first and the second derivative[2] of the estimated risk $\hat{R}(\mathbf{c})$ could be obtained by doing derivation of $\hat{R}(\mathbf{c})$ on $\operatorname{div}(f_{ens}, \mathbf{x}^*)$,, i.e.,

$$\hat{R}' = 4\left(\frac{8\delta}{1-2\varepsilon}\right)^2 \frac{1}{(\lambda - 2\operatorname{div}(f_{ens}, \mathbf{x}^*))^3} \cdot \log_2\left(8|S|\left(\frac{(1-2\varepsilon)(\lambda - 2\operatorname{div}(f_{ens}, \mathbf{x}^*))}{8\delta}\right)^2\right), \tag{32}$$

$$\hat{R}'' = \frac{8}{\ln 2}\left(\frac{8\delta}{1-2\varepsilon}\right)^2 \frac{1}{(\lambda - 2\operatorname{div}(f_{ens}, \mathbf{x}^*))^4} \cdot \left(3\ln\left(8|S|\left(\frac{(1-2\varepsilon)(\lambda - 2\operatorname{div}(f_{ens}, \mathbf{x}^*))}{8\delta}\right)^2\right) - 2\right). \tag{33}$$

We are already aware that $\operatorname{div}(f_{ens}, \mathbf{x}^*) \in (-\frac{1}{2}, \frac{1}{2})$. Thus, the graph of the upper bound of the generalization error $\hat{R}(\mathbf{c})$ is symmetrical about the vertical axis, as shown in Figure 2. And then we take $\operatorname{div}(f_{ens}, \mathbf{x}^*) \in (0, \frac{1}{2})$ for example to analyze the effect of diversity on the generalization error in the ensemble.

As the relationship between the proposed diversity $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ and the generalization error $\hat{R}(\mathbf{c})$ in the ensemble is Eq. (31), the monotonicity and concavity of function could be analyzed based on its first derivative and second derivative. On the positive abscissa axis, let q_1, q_3 be the left endpoint and the right endpoint of the range, respectively. The function inflexion q_2 is found by making $\hat{R}'(\mathbf{c}) = 0$, also remarked as q_4 . Thus the monotone increasing interval is (q_1, q_2) , and the monotone diminishing interval is (q_2, q_3) . Then the stagnation point of the first derivative q_5 is found by making $\hat{R}''(\mathbf{c}) = 0$, to analyze the concavity of the function $\hat{R}(\mathbf{c})$; the stagnation point of the second derivative q_6 is found by making $\hat{R}'''(\mathbf{c}) = 0$, to analyze the concavity of the function $\hat{R}'(\mathbf{c})$. Their values are

$$q_1 = \varepsilon\,, \quad q_3 = \frac{1}{2}\left(1 - \frac{\varepsilon}{1-2\varepsilon}\right), \tag{34a}$$

$$q_4 = q_2 = \frac{1}{2}\left(1 - \frac{\delta}{1-2\varepsilon}\sqrt{\frac{8}{|S|}}\right), \tag{34b}$$

$$q_5 = \frac{1}{2}\left(1 - \frac{\delta}{1-2\varepsilon}\sqrt{\frac{8}{|S|}e^{\frac{2}{3}}}\right), \tag{34c}$$

$$q_6 = \frac{1}{2}\left(1 - \frac{\delta}{1-2\varepsilon}\sqrt{\frac{8}{|S|}e^{\frac{7}{6}}}\right), \tag{34d}$$

where an implied condition exists, i.e.,

$$\varepsilon \leqslant \frac{\delta}{1-2\varepsilon}\sqrt{\frac{8}{|S|}} \leqslant 1 - 2\varepsilon\,. \tag{35}$$

Note that $q_1 < q_2 < q_3$ and $q_6 < q_5 < q_4 = q_2$. Then the domain of diversity $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ could be divided into eight intervals, as shown in Table I. Up to now, we could analyze

TABLE I
MONOTONE INTERVALS. THE FIRST COLUMN IS DIVERSITY $\operatorname{div}(f_{ens}, \mathbf{x}^*)$. THE SECOND AND THE THIRD COLUMNS ARE THE ESTIMATED RISK $\hat{R}(\mathbf{c})$ AND ITS FIRST DERIVATIVE, RESPECTIVELY. NOTE THAT THE ESTIMATED RISK REFLECTS THE GENERALIZATION ERROR $R(\mathbf{c})$ OF THE ENSEMBLE f_{ens}. THE FOURTH AND THE FIFTH COLUMNS ARE THE VARIATION OF THE ESTIMATED RISK $\hat{R}(\mathbf{c})$ AND THAT OF ITS FIRST DERIVATIVE, RESPECTIVELY.

$\operatorname{div}(f_{ens}, \mathbf{x}^*)$	$\hat{R}(\mathbf{c})$	$\hat{R}'(\mathbf{c})$	$\Delta\hat{R}$	$\Delta\hat{R}'$
$(-q_3, -q_2)$	$\nearrow$ convex	$\searrow$ concave	smaller	larger
$(-q_2, -q_5)$	$\searrow$ convex	$\searrow$ concave	smaller	larger
$(-q_5, -q_6)$	$\searrow$ concave	$\nearrow$ concave	larger	larger
$(-q_6, -q_1)$	$\searrow$ concave	$\nearrow$ convex	larger	smaller
(q_1, q_6)	$\nearrow$ concave	$\nearrow$ concave	larger	larger
(q_6, q_5)	$\nearrow$ concave	$\nearrow$ convex	larger	smaller
(q_5, q_2)	$\nearrow$ convex	$\searrow$ convex	smaller	smaller
(q_2, q_3)	$\searrow$ convex	$\searrow$ convex	smaller	smaller

[2]The first derivative of a function $f(x)$, written as $f'(x)$, is the slope of the tangent line to the function at the point x. To put this in non-graphical terms, the first derivative tells us whether a function is increasing or decreasing, and by how much it is increasing or decreasing. The second derivative of a function is the derivative of the derivative of that function, written as $f''(x)$. While the first derivative can tell us if the function is increasing or decreasing, the second derivative tells us if the first derivative is increasing or decreasing [50].

in detail the effect of diversity on generalization error within different intervals.

(1) When $\operatorname{div}(f_{ens}, \mathbf{x}^*) = 0$ in this paper, we only consider two cases: "all individual classifiers give the right result of the instance $\mathbf{x}^*$'s label;" "all individual classifiers give the wrong result of the instance $\mathbf{x}^*$'s label." We do not consider the situation of a tie.

(2) In real situations, $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ is negative in most cases which means "the ensemble classifier has misclassified the instance that corresponds to the selected $\operatorname{div}(f_{ens}, \mathbf{x}^*)$." Notice that in the finite experiments, when the value of $|S|$ satisfies Eq. (35), it is rare that $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ is larger than zero. If $\operatorname{div}(f_{ens}, \mathbf{x}^*) < 0$, increasing diversity will improve the ensemble's performance, which is the reason why we should increase diversity in that case.

(3) When $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ is located in the positive abscissa axis, the ensemble classifies the corresponding instance correctly. In the range near zero, individual classifiers that classify the instance correctly take great advantage, and the closer $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ is to zero, the greater their advantage is. In the range near $1/2$, individual classifiers that classify the instance correctly take a small advantage, and the closer $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ is to $1/2$, the smaller their advantage is. In this case, diversity needs to be decreased to increase the generalization performance of the ensemble, with the basic idea of keeping the ensemble classifying the corresponding instance correctly. In summary, increasing diversity will reduce the advantage of the individuals that classify the instance correctly, which should be avoided.

(4) When $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ is located in the negative axis, the ensemble classifies the instance incorrectly. Similar results could be analogized.

Based on these analyses, the relationship between the proposed diversity and generalization error varies when diversity lays in different ranges. Generally speaking, there are two specific ranges that need to be paid attention to particularly, i.e., $(-q_5, -q_6)$ and $(-q_6, -q_1)$, according to Table I. In these two ranges, more diversity would lead to better generalization, which is why it is necessary to increase $\operatorname{div}(f_{ens}, \mathbf{x}^*)$ to get a more effective ensemble classifier. However, in other ranges, more diversity will not lead to better generalization, so the ensemble could remain the same to continue the learning process.

D. Ensemble Pruning based on Diversity

Inspired by the relationship between the proposed diversity and generalization of classification ensembles, we propose an ensemble pruning method, named as "*Ensemble Pruning based on Diversity (EPBD)*," presented in Algorithm 1. It aims to prune an ensemble with rare performance degradation, with the basic idea of utilizing diversity and accuracy simultaneously.

The inputs of this method are: (i) a training set S of valid data instances, (ii) a set F of trained individual classifiers, i.e., the original set of classifiers to constitute an ensemble classifier in Eq. (1), (iii) the ratio α which means the individual classifiers in the pruned sub-ensemble are chosen in a ratio of $\alpha(\%)$ from the original ensemble, and (iv) the trade-off parameter β is used to balance accuracy and diversity. Notice that the coefficient c corresponding to the individual f is set to $1/|F|$ for simplicity here. The output of this method is the set of classifiers H composing the pruned sub-ensemble after pruning, which is initially set to $\emptyset$. This algorithm terminates when the number of H reaches the number of a pruned sub-ensemble classifier, i.e., $\alpha|F|$, or when it cannot pick up an individual classifier in F that satisfies the corresponding condition (i.e., all individual classifiers in F misclassify the corresponding instance).

Algorithm 1 Ensemble Pruning based on Diversity, *EPBD*

Input: A training set $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{|S|}, y_{|S|})\}$,
an original ensemble $F = \{f_1, ..., f_{|F|}\}$,
the threshold μ, and the trade-off parameter β.
Output: The pruned sub-ensemble $H \subset F$.
1: $H = \emptyset$.
2: **repeat**
3: Search for the specific data instance $(\mathbf{x}, y)$ which satisfies the search criterion Eq. (30).
4: Sort classifiers in F that classify this instance correctly in ascending order according to generalization performance.
5: Pick out the first one $f(\cdot)$ in the previous step, add it to H, and remove it from F.
6: **until** The termination condition is met.

The objective of this method is to obtain a pruned sub-ensemble whose performance approaches as optimal as possible. As shown in Algorithm 1, it firstly picks out an individual classifier whose diversity and accuracy are as high as possible, raising diversity in line 3 and increasing accuracy in line 4. Specifically, the sorting principle we use in line 4 is mainly based on accuracy plus β times $\operatorname{div}(f_{ens}, \mathbf{x})$ in Eq. (30). Then it puts the selected individual classifier into the candidate classifier set H, as shown in line 5. Subsequently, it follows a cycle to select the classifiers in F repeatedly until the termination conditions are met.

E. Complexity Analysis of EPBD

In this subsection, we give the complexity analysis of the proposed *EPBD*. According to the Algorithm 1, the computational complexity of *EPBD* is analyzed as follows:

- Firstly, the complexity of line 3 is $\mathcal{O}(jm + m\log m)$ where $j = n-i+1$, and n, m are the number of individual classifiers and instances, respectively, when $i \in \{1, ..., k\}$. Note that k is the size of the pruned sub-ensemble.
- Secondly, the complexity of line 4 is $\mathcal{O}(j \log j)$ where $j = n - i + 1$ when $i \in \{1, ..., k\}$.
- Thirdly, the complexity of line 5 is $\mathcal{O}(1)$.

Therefore, the overall computational complexity of *EPBD* is $\mathcal{O}\big(\sum_{j=n-k+1}^{n}(jm + m\log m + j\log j)\big)$ which is $\mathcal{O}\big(mk(n - \frac{k-1}{2}) + mk\log m + \sum_{i=n-k+1}^{n}(i\log i)\big)$.

IV. Experiments

In this section, we elaborate our experiments to evaluate the proposed relationship between the proposed diversity and generalization performance of ensemble classifiers and the proposed algorithm. The data sets we use includes one image

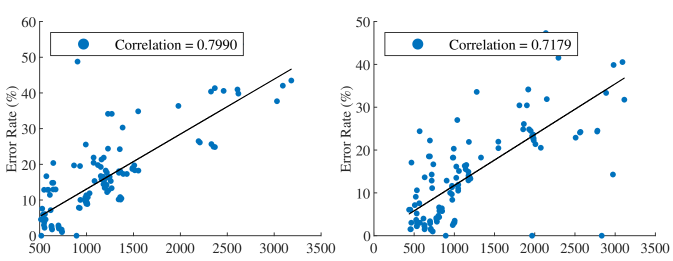
TABLE II
COMPARISON OF THE STATE-OF-THE-ART PRUNING METHODS WITH *EPBD* USING BAGGING TO PRODUCE AN ENSEMBLE WITH LMS AS INDIVIDUAL CLASSIFIERS. THE COLUMN NAMED AS "ENSEM" IS THE CORRESPONDING RESULTS FOR THE ENTIRE ENSEMBLE WITHOUT PRUNING. THE BEST ACCURACY WITH LOWER STANDARD DEVIATION IS INDICATED WITH BOLD FONTS FOR EACH DATA SET (ROW).

Data set	Ensem	ES	KL	KP	OO	DREP	SEP	PEP	GMA	LCS	*EPBD*
card	75.77±6.14	75.18±6.02	77.81±5.90	74.16±5.13	77.96±6.31	**79.27±5.60**	75.04±6.89	77.96±4.83	74.31±6.66‡	73.58±5.71	77.52±6.64
heart	82.59±4.06	82.96±4.01	82.96±4.97	83.33±2.27	82.96±1.55	82.96±3.04	**83.70±3.56**	82.22±2.11	81.48±2.93	81.85±3.31	83.33±2.93
ringnorm	77.38±0.88	77.09±0.66	77.12±0.87	77.34±0.96	77.36±0.83	76.86±0.65‡	**77.57±0.97**	77.35±0.90	77.27±0.70	77.19±0.92	77.46±0.75
spam	89.79±1.32	89.73±1.01	89.75±1.26	89.40±1.05	89.36±1.06	**90.03±0.90**	89.68±1.24	90.01±1.13	89.62±1.21	89.51±1.17	89.58±1.02
waveform	88.27±1.14	88.31±1.13	88.39±1.15	88.33±1.23	**88.45±1.16**	88.31±1.16	88.29±1.20	88.37±1.08	88.29±1.07	88.29±1.06	88.33±1.10
wisconsin	96.74±1.24	96.74±1.12	96.74±1.44	96.59±1.35	96.59±1.86	96.44±1.10	96.59±0.99	96.74±1.35	96.74±1.12	**96.74±0.84**	**96.74±0.84**
credit	75.99±0.65‡	75.18±1.00‡	69.96±3.46‡	67.62±14.58	77.27±0.36	**77.81±0.11**†	75.93±1.17‡	76.95±1.19	74.90±3.56	76.65±0.76	77.45±0.25
landsat	96.28±0.36	96.30±0.32	96.25±0.36	96.30±0.30	96.31±0.36	96.13±0.32‡	96.28±0.27	96.25±0.31	96.27±0.29	**96.31±0.28**	96.30±0.42
ecoli	91.21±3.92	88.79±3.65	87.88±4.15‡	81.82±7.03‡	89.39±6.52	**92.73±4.20**	90.61±3.92	91.52±4.49	84.55±8.73	87.58±8.40	90.91±4.42
mammo graphic	82.40±2.03	81.67±3.11	81.98±2.38	81.15±2.93	81.77±1.91	82.60±2.13	81.77±2.83	**82.60±1.41**	81.15±1.70	81.98±1.25	81.88±2.13
segmentation	28.18±0.36‡	91.00±2.27	91.13±2.31	91.17±1.84	**91.39±2.01**	89.87±3.91	91.13±2.31	91.26±2.23	90.95±2.04	90.95±2.11‡	91.30±2.09
sensor (2d)	62.62±1.42	62.55±1.37‡	62.53±1.72	62.44±1.54	62.99±1.27	61.58±1.36‡	62.75±1.55	63.06±1.64	62.55±1.79	62.57±1.37	**63.13±1.17**
sensor (4d)	62.99±1.79‡	62.91±1.96‡	62.90±2.10‡	62.86±1.84‡	63.56±2.10	62.40±1.96‡	63.08±1.87‡	63.43±1.92	63.12±1.96‡	63.06±2.08‡	**63.59±2.14**
waveform	85.65±1.57	85.53±1.59	85.35±1.66	85.65±1.84	85.55±1.15	84.84±2.03	85.53±1.47	**85.87±1.09**	85.45±1.46	85.19±1.26	85.83±1.32
waveform_noise	85.53±1.17	85.51±1.67	85.15±1.10‡	85.39±1.27	85.11±0.92	84.60±1.41	85.13±1.36	85.41±1.51	85.57±1.14	85.39±1.34	**85.63±1.15**
t-test (W/T/L)	3/17/0	4/16/0	4/16/0	2/18/0	0/20/0	4/15/1	2/18/0	0/20/0	3/17/0	3/17/0	—
Average Rank	5.53	6.60	6.73	7.50	4.80	6.97	6.00	3.77	7.63	7.20	3.27

[1] The reported results are the average test accuracy (%) of each method and the corresponding standard deviation under 5-fold cross-validation on each data set.
[2] By two-tailed paired t-test at 5% significance level, ‡ and † denote that the performance of *EPBD* is superior to and inferior to that of the comparative method, respectively.
[3] The last two rows show the results of t-test and average rank, respectively. The "W/T/L" in t-test indicates that *EPBD* is superior to, not significantly different from, or inferior to the corresponding comparative methods. The average rank is calculated according to the Friedman test [51].

TABLE III
COMPARISON OF THE SPACE COST BETWEEN *EPBD* AND OTHER METHODS THAN CANNOT FIX THE SIZE OF PRUNED SUB-ENSEMBLES, USING BAGGING TO CONSTRUCT CLASSIFICATION ENSEMBLES. NOTE THAT *EPBD* COULD GUARANTEE THAT THE SIZE OF THE PRUNED SUB-ENSEMBLE IS NO MORE THAN THE UP LIMIT OF THAT, BASED ON THE PRUNING RATE. THE BEST PERFORMANCE (HIGHER ACCURACY WITH LOWER STANDARD DEVIATION) WITH ITS CORRESPONDING SIZE OF THE PRUNED SUB-ENSEMBLE IS INDICATED WITH BOLD FONTS FOR EACH DATA SET (ROW).

Data set	Test Accuracy (%) / Size of Pruned Sub-Ensembles											
	OO		DREP		SEP		PEP		LCS		*EPBD*	
landsat	**97.60±0.23**	**9.80±1.30**	95.89±0.43‡	1.00±0.00	97.51±0.47	9.00±2.35	97.51±0.10	11.00±0.71	97.42±0.52	9.00±0.00	97.53±0.19	6.80±1.64
mammo graphic	79.90±1.75‡	9.60±1.82	80.31±1.62	1.20±0.45	79.58±1.49	10.00±1.22	80.10±1.82	7.40±1.67	80.31±1.96	9.00±0.00	**80.52±2.07**	**6.00±1.00**
sensor (24d)	90.34±1.03	10.40±1.34	90.14±0.80	1.00±0.00	90.25±0.84	9.00±3.32	90.34±1.13	9.00±2.00	90.03±0.67‡	9.00±0.00	**90.50±0.85**	**7.60±0.89**
waveform	88.07±0.56	10.20±1.64	87.97±0.35	1.00±0.00	88.13±0.80	9.40±2.61	88.07±0.62	8.60±2.07	88.01±0.95	9.00±0.00	**88.19±0.72**	**6.60±1.14**
ringnorm	77.36±0.83	11.40±1.14	76.86±0.65‡	1.00±0.00	**77.57±0.97**	**6.80±2.28**	77.35±0.90	6.80±1.79	77.19±0.92	9.00±0.00	77.46±0.75	6.40±0.55
landsat	96.31±0.36	10.60±1.14	96.13±0.32‡	1.00±0.00	96.28±0.27	9.40±1.34	96.25±0.31	6.60±1.95	**96.31±0.28**	**9.00±0.00**	96.30±0.42	6.40±1.34
madelon	54.35±2.53	11.00±2.00	53.92±2.42	1.00±0.00	54.50±2.18‡	9.00±2.55	54.46±2.12‡	11.80±1.64	55.04±2.82	9.20±0.45	**55.15±2.49**	**8.40±0.89**
credit	74.23±0.61	10.80±1.10	70.91±0.22‡	1.00±0.00	73.79±0.63	8.60±1.67	74.50±0.30	21.00±0.00	73.84±0.40	9.00±0.00	**74.76±0.65**	**5.80±0.45**
t-test (W/T/L)	1/7/0		4/4/0		1/7/0		1/7/0		1/7/0		—	
Average Rank	3.06		5.44		3.56		3.56		3.88		1.50	



(a) (b)

Fig. 3. Relationship of error rate and estimated risk in Eq. (25) calculated based on diversity in Eq. (11) for binary classification. (a) Using bagging with NBs as individual classifiers. (b) Using bagging with LMs as individual classifiers.

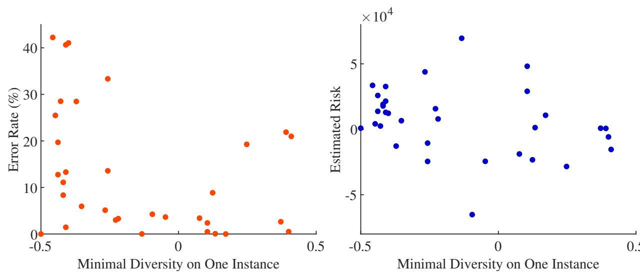
data set with $12,500$ pictures (Dogs vs Cats[3]) and 28 data sets from UCI repository[4] [52]. Standard 5-fold cross-validation is used in these experiments, i.e., in each iteration, the entire data set is split into two parts, with 80% as the training set and 20% as the test set. Bagging [22] and AdaBoost [25], [53] are used to constitute an ensemble classifier on various kinds of individual classifiers including decision trees (DT), naive Bayesian (NB) classifiers, k-nearest neighbors (KNN) classifiers, linear model (LM) classifiers, and linear SVMs (LSVM). We evaluate the proposed relationship between the proposed diversity and ensemble performance using scatter diagrams with their correlation. Besides, to evaluate our proposed pruning method *EPBD*, the baselines we considered are a variety of ranking-based methods, namely Early Stopping (ES), KL-divergence pruning (KL), Kappa pruning (KP) [54], orientation ordering pruning (OO) [55], and diversity regularized ensemble pruning (DREP) [29] as well as optimization-based methods, namely single-objective ensemble pruning (SEP), and Pareto ensem-

[3]http://www.kaggle.com/c/dogs-vs-cats

[4]http://archive.ics.uci.edu/ml/datasets.html

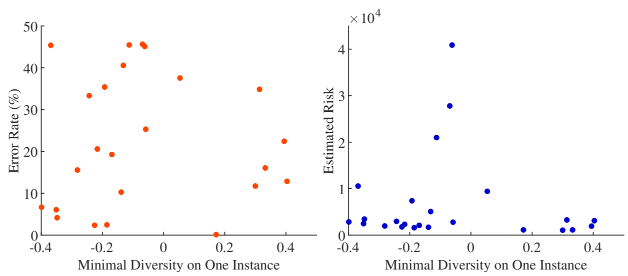
TABLE IV
COMPARISON OF THE SPACE COST BETWEEN *EPBD* AND OTHER METHODS THAN CANNOT FIX THE SIZE OF PRUNED SUB-ENSEMBLES, USING ADABOOST TO CONSTRUCT CLASSIFICATION ENSEMBLES. NOTE THAT *EPBD* COULD GUARANTEE THAT THE SIZE OF THE PRUNED SUB-ENSEMBLE IS NO MORE THAN THE UP LIMIT OF THAT, BASED ON THE PRUNING RATE. THE BEST PERFORMANCE (HIGHER ACCURACY WITH LOWER STANDARD DEVIATION) WITH ITS CORRESPONDING SIZE OF THE PRUNED SUB-ENSEMBLE IS INDICATED WITH BOLD FONTS FOR EACH DATA SET (ROW).

Data set	Test Accuracy (%) / Size of Pruned Sub-Ensembles											
	OO		DREP		SEP		PEP		LCS		*EPBD*	
madelon	60.46±2.49	9.60±1.14	**64.73±2.01**	**1.00±0.00**	55.92±1.76	9.40±0.89	60.08±2.39	12.60±3.29	63.73±3.88	9.00±0.00	61.92±4.14	9.00±0.00
waveform_noise	83.70±1.88	8.80±1.30	79.70±1.82‡	1.00±0.00	82.64±1.01	10.80±1.92	83.54±0.93	10.40±2.61	**84.68±1.17**	**9.00±0.00**	84.32±1.29	9.00±0.00
spam	83.87±2.97	5.20±1.30	**85.40±2.90**	**1.00±0.00**	80.28±5.43	10.60±0.89	71.49±29.28	4.60±2.97	29.36±9.07‡	9.20±0.45	83.94±2.85	9.00±0.00
iono	92.29±2.78	8.00±0.71	**92.86±2.47**	**1.00±0.00**	91.14±5.19	10.80±2.28	88.86±4.33	3.60±2.61	32.86±28.12‡	9.20±0.45	**92.86±2.47**	**6.40±2.07**
waveform	**88.07±1.12**	**1.80±0.45**	**88.07±1.12**	**1.00±0.00**	84.20±9.24	8.80±3.03	84.62±8.31	3.00±1.41	60.70±15.69‡	9.00±0.00	**88.07±1.12**	**7.20±0.84**
ringnorm	**76.27±1.61**	**9.20±0.45**	**76.27±1.61**	**1.00±0.00**	63.07±2.40‡	9.40±2.61	**76.27±1.61**	**2.00±0.00**	72.32±8.79	9.00±0.00	**76.27±1.61**	**9.00±0.00**
sonar	74.15±9.85	10.20±1.48	70.73±6.90‡	1.00±0.00	73.66±7.19	8.80±1.79	77.07±7.24	10.80±1.10	76.59±4.43	9.00±0.00	**77.56±6.07**	**8.80±0.45**
wilt	97.46±0.71	11.80±0.84	97.48±0.88	1.40±0.89	97.08±0.38‡	8.00±3.16	97.46±0.57	10.00±1.41	97.54±0.62	9.40±0.55	**97.58±0.62**	**9.00±0.00**
ecoli	63.33±5.29‡	10.40±1.14	64.24±8.41	1.00±0.00	62.73±6.21	11.40±1.52	67.27±8.34	10.00±1.00	64.55±7.55	9.00±0.00	**67.27±6.83**	**9.00±0.00**
t-test (W/T/L)	1/8/0		2/7/0		2/7/0		0/9/0		3/6/0		—	
Average Rank	3.44		3.00		5.22		3.72		3.78		1.83	



(a) (b)

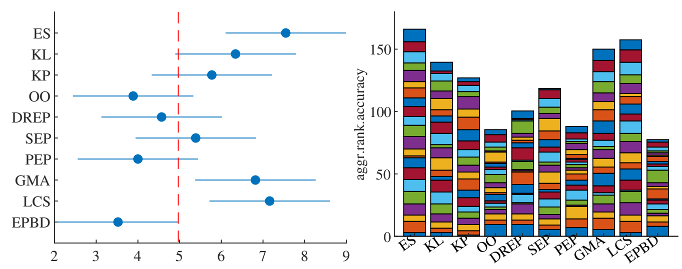
Fig. 4. Relationship of diversity and ensemble performance in Theorem 4, using bagging with DTs as individual classifiers for binary classification. (a) Relationship between the proposed diversity and error rate of the ensemble. (b) Relationship between the proposed diversity and the estimated risk in Eq. (25).



(a) (b)

Fig. 5. Relationship of diversity and ensemble performance in Theorem 4, using AdaBoost with LMs as individual classifiers for binary classification. (a) Relationship between the proposed diversity and error rate of the ensemble. (b) Relationship between the proposed diversity and the estimated risk in Eq. (25).

ble pruning (PEP) [56]. Moreover, we adopt two methods from diversity maximization via composable coresets [57] and change them slightly to make them suitable for pruning problems, namely the Gonzalez's algorithm (GMA) and local search algorithm (LCS). An ensemble is trained and pruned on the training set and then tested on the test set. It is worth mentioning that several methods (such as OO, DREP, SEP, PEP, and LCS) cannot fix the number of classifiers after ensemble pruning, while others could do that by giving a



(a) (b)

Fig. 6. Comparison of the state-of-the-art methods with *EPBD* on the test accuracy. (a) Friedman test chart (non-overlapping means significant difference) [51]. (b) The aggregated rank for each method (the smaller the better) [56].

pruning rate that is the up limit of the percentage of those discarded individual classifiers in the original ensemble. Those methods that cannot fix the size of the pruned sub-ensemble might lead to increase or reduce the size of pruned sub-ensembles and affect their space cost.

A. Validating the Relationship Between the Proposed Diversity and Ensemble Performance

In this subsection, experiments are conducted to verify the proposed relationship between the proposed diversity and generalization performance of ensemble classifiers (Theorem 4). Before that, we need to verify the ensemble performance's relationship with the estimated risk. The experimental results are reported in Figure 3, including different individual classifiers. In each experiment, estimated risk was calculated by Eq. (25), based on the proposed diversity by Eq. (11). The correlation between estimated risk and real error rate was calculated, as annotations shown in Figures 3(a)–3(b), presenting a high level of correlation between them. Besides, experimental results reported in Figures 4–5 indicate that the relationship between the proposed diversity and ensemble performance in Theorem 4 is faithful. Along with the increase of diversity, the ensemble performance would be on an "upward-downward-upward-downward" trend in Figures 4(a) and 5(a), which coincides with the analyses in Section III-C. Therefore, we

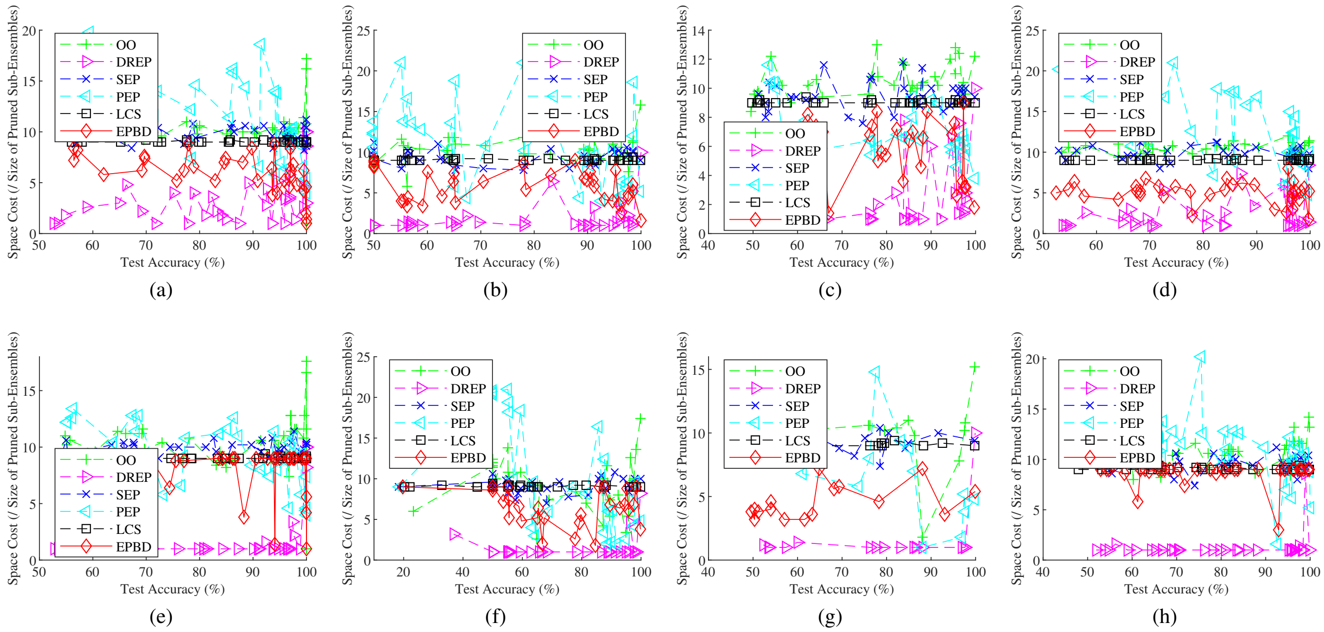


Fig. 7. Comparison of the space cost between *EPBD* and other methods that cannot fix the size of pruned sub-ensembles. Although *EPBD* cannot fix the number of individual classifiers in the pruned sub-ensemble, it could guarantee that the size of the pruned sub-ensemble is no more than the up limit of that, based on the pruning rate. (a–d) Using bagging with DTs, SVMs, LSVMs, and KNNs as individual classifiers, respectively. (e–h) Using AdaBoost with DTs, SVMs, LSVMs, and KNNs as individual classifiers, respectively.

could utilize the proposed relationship between the proposed diversity and generalization error to guide our understanding of diversity, and this relationship is a good start to dig the real role of diversity in ensemble learning.

B. Comparison Between EPBD *and the State-of-the-art Ensemble Pruning Methods*

In this subsection, we compare the performance of various ensemble pruning methods, including ES, KL, KP, OO, DREP, SEP, PEP, GMA, and LCS with the proposed *EPBD* method. Experimental results reported in Table II contain the average test accuracy of each method and the corresponding standard deviation under 5-fold cross-validation on each data set. Each row (data set) in Table II compares the classification accuracy using bagging with the same type of individual classifiers, indicating results with higher accuracy and lower standard deviation by bold fonts. Besides, the significance of the difference in the accuracy performance between two methods is examined by two-tailed paired t-test at 5% significance level to tell if two ensemble pruning methods have significantly different results. Two methods end in a tie if there is no significant statistical difference; otherwise, one with higher values of accuracy will win. The performance of each method is reported in the last row of Table II, compared with *EPBD* in terms of the number of data sets that *EPBD* has won, tied, or lost, respectively. We may notice that sometimes OO or PEP might achieve the best results instead of our *EPBD*. However, both of them cannot fix the number of classifiers after ensemble pruning like *EPBD* does, leading to keep more individual classifiers than we desire in most situations. Therefore, sometimes they may achieve better results than *EPBD* yet without significant difference. Meanwhile, it could be inferred that *EPBD* could achieve competitive results even though it only keeps fewer individual classifiers, which means that the principle we used to guide the pruning process is effectual and that utilizing our proposed relationship is reasonable. Figure 6(a) shows that *EPBD* could achieve competitive results as OO and PEP, and *EPBD* has significant superiority over other compared pruning methods. Figure 6(b) presents the aggregated rank for each method, describing the similar conclusions to Figure 6(a).

C. Comparison of Space Cost Between EPBD *and Methods that Cannot Fix the Size of Pruned Sub-Ensembles*

In this subsection, we compare the accuracy and the space cost between *EPBD* with methods that cannot fix the size of pruned sub-ensembles, such as OO, DREP, SEP, PEP, and LCS. Experimental results are reported in Tables III–IV and Figure 7. As shown in Figure 7, DREP only keeps one of the individual classifiers that is not an ensemble at all in most cases; OO, SEP, and PEP usually generate pruned sub-ensembles with a larger size than the up limit of that; the size of pruned sub-ensembles generated by LCS is relatively steady around the up limit of that. Besides, although *EPBD* cannot fix the number of individual classifiers in the pruned sub-ensemble, it could guarantee that the size of the pruned sub-ensemble is no more than the up limit of that, based on the pruning rate. Therefore, we could conclude that *EPBD* could run as expected to generate a comparable ensemble performance with a smaller size to the original ensemble.

V. Conclusion

This paper has investigated and utilized diversity in classification ensembles, and it made the following contributions to the ensemble learning community. First of all, inspired by the regression ensembles, this paper proposed the measure of diversity utilizing the error decomposition for classifier ensembles, which broke the error of classification ensembles into two terms: the accuracy and the diversity terms. The empirical results have confirmed that "div" is a distinct diversity measure. Secondly, we have theoretically investigated the relationship between the proposed diversity and generalization of the ensemble through the bound between margin and generalization. These theoretical analyses solve an open question in ensemble area to some extent, i.e., when does diversity help the generalization of ensembles? Based on the analyses, the relationship between the proposed diversity and generalization error varies when diversity is in different ranges. In some ranges, more diversity could lead to better generalization, while in other ranges, more diversity is not beneficial to the generalization based on the bound analyses. Thirdly, in order to validate and employ the relationship of diversity to improve the performance of ensembles, we have proposed *EPBD* that can prune an ensemble without much performance degradation. Although this generalization bound could be loose, our work provides a direction and one theoretical attempt to reveal the impacts of the diversity to the generalization. Our future work is to generalize the proposed methodology to the multi-classification problems.

References

[1] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC press, 2012.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[3] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012, pp. 1290–1297.

[4] X. Zhou, L. Xie, P. Zhang, and Y. Zhang, "An ensemble of deep neural networks for object tracking," in *ICIP*, Oct 2014, pp. 843–847.

[5] H. Ykhlef and D. Bouchaffra, "An efficient ensemble pruning approach based on simple coalitional games," *Inform Fusion*, vol. 34, pp. 28–42, 2017.

[6] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowl Inf Syst*, vol. 14, no. 1, pp. 1–37, 2008.

[7] L. Rokach, "Ensemble-based classifiers," *Artif Intell Rev*, vol. 33, no. 1-2, pp. 1–39, 2010.

[8] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, "Ensemble-based classifiers," in *Multilabel Classification*. Springer, 2016, pp. 101–113.

[9] S. Gu, R. Cheng, and Y. Jin, "Multi-objective ensemble generation," *Wires Data Min Knowl*, vol. 5, no. 5, pp. 234–245, 2015.

[10] T. Dietterich, "Ensemble methods in machine learning," in *MCS*, vol. 1857, 2000, pp. 1–15.

[11] Z. Lu, X. Wu, X. Zhu, and J. Bongard, "Ensemble pruning via individual contribution ordering," in *SIGKDD*. ACM, 2010, pp. 871–880.

[12] E. Tang, P. Suganthan, and X. Yao, "An analysis of diversity measures," *Mach Learn*, vol. 65, no. 1, pp. 247–271, jul 2006.

[13] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach Learn*, vol. 51, no. 2, pp. 181–207, 2003.

[14] G. Brown, "An information theoretic perspective on multiple classifier systems," in *MCS*, 2009, pp. 344–353.

[15] Z.-H. Zhou and N. Li, "Multi-information ensemble diversity," in *MCS*, 2010, pp. 134–144.

[16] Y. Yu, Y.-F. Li, and Z.-H. Zhou, "Diversity regularized machine," in *IJCAI*, 2011, p. 1603.

[17] Z. Jiang, H. Liu, B. Fu, and Z. Wu, "Generalized ambiguity decompositions for classification with applications in active learning and unsupervised ensemble pruning." in *AAAI*, 2017, pp. 2073–2079.

[18] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *NIPS*, 1995, pp. 231–238.

[19] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput*, vol. 4, no. 1, pp. 1–58, 1992.

[20] R. Herbrich and T. Graepel, "A pac-bayesian margin bound for linear classifiers: Why svms work," in *NIPS*, 2001, pp. 224–230.

[21] ——, "A pac-bayesian margin bound for linear classifiers," *IEEE T Inform Theory*, vol. 48, no. 12, pp. 3140–3150, 2002.

[22] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, 1996.

[23] Y. Freund and R. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *EuroCOLT*. Springer, 1995, pp. 23–37.

[24] ——, "A decision-theoretic generalization of on-line learning and an application to boosting," *J Comput Syst Sci*, vol. 55, no. 1, pp. 119–139, 1997.

[25] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *ICML*, vol. 96. Bari, Italy, 1996, pp. 148–156.

[26] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[27] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE T Pattern Anal*, vol. 28, no. 10, pp. 1619–1630, 2006.

[28] A. Chandra and X. Yao, "Divace: Diverse and accurate ensemble learning algorithm," in *IDEAL*. Springer, 2004, pp. 619–625.

[29] N. Li, Y. Yu, and Z.-H. Zhou, "Diversity regularized ensemble pruning," in *ECML PKDD*, 2012, pp. 330–345.

[30] F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems," in *MCS*. Springer, 2001, pp. 78–87.

[31] S. Gu and Y. Jin, "Generating diverse and accurate classifier ensembles using multi-objective optimization," in *MCDM*. IEEE, 2014, pp. 9–15.

[32] G. U. Yule, "On the association of attributes in statistics: with illustrations from the material of the childhood society, &c," *Phil. Trans. R. Soc. Lond. A*, vol. 194, no. 252-261, pp. 257–319, 1900.

[33] J. Cohen, "A coefficient of agreement for nominal scales," *Educ Psychol Meas*, vol. 20, no. 1, pp. 37–46, 1960.

[34] D. B. Skalak *et al.*, "The sources of increased accuracy for two proposed boosting algorithms," in *AAAI*, vol. 1129, 1996, p. 1133.

[35] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE T Pattern Anal*, vol. 20, no. 8, pp. 832–844, 1998.

[36] P. Sneath and R. Sokal, *Numerical taxonomy: The principles and practice of numerical classification.* WH Freeman, 1973.

[37] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image Vision Comput*, vol. 19, no. 9, pp. 699–707, 2001.

[38] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[39] R. Kohavi, D. H. Wolpert *et al.*, "Bias plus variance decomposition for zero-one loss functions," in *ICML*, vol. 96, 1996, pp. 275–83.

[40] P. Cunningham and J. Carney, "Diversity versus quality in classification ensembles based on feature selection," in *ECML*. Springer, 2000, pp. 109–116.

[41] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Inform Fusion*, vol. 3, no. 2, pp. 135–148, 2002.

[42] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE T Pattern Anal*, vol. 12, no. 10, pp. 993–1001, 1990.

[43] D. Partridge and W. Krzanowski, "Software diversity: practical statistics for its measurement and exploitation," *Inform Software Tech*, vol. 39, no. 10, pp. 707–717, 1997.

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

[45] H. Chen and X. Yao, "Multiobjective neural network ensembles based on regularized negative correlation learning," *IEEE T Knowl Data En*, vol. 22, no. 12, pp. 1738–1751, 2010.

[46] G. Zenobi and P. Cunningham, "Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error," in *ECML*. Springer, 2001, pp. 576–587.

[47] L. G. Valiant, "A theory of the learnable," *Commun ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[48] G. Brown, J. L. Wyatt, and P. Tiňo, "Managing diversity in regression ensembles," *J Mach Learn Res*, vol. 6, no. Sep, pp. 1621–1650, 2005.

[49] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Comput Surv*, vol. 45, no. 1, p. 10, 2012.

[50] D. M. Department, "The first and second derivatives," Website, accessed on October 26, 2019. [Online]. Available: https://math.dartmouth.edu/opencalc2/cole/lecture8.pdf

[51] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J Mach Learn Res*, vol. 7, no. Jan, pp. 1–30, 2006.
[52] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml
[53] Y. Freund, "Boosting a weak learning algorithm by majority," *Inform Comput*, vol. 121, no. 2, pp. 256–285, 1995.
[54] D. Margineantu and T. Dietterich, "Pruning adaptive boosting," in *ICML*, vol. 97, 1997, pp. 211–218.
[55] Martínez-Muñoz and Suárez, "Pruning in ordered bagging ensembles," in *ICML*, 2006, pp. 609–616.
[56] C. Qian, Y. Yu, and Z.-H. Zhou, "Pareto ensemble pruning," in *AAAI*, 2015, pp. 2935–2941.
[57] S. Aghamolaei, M. Farhadi, and H. Zarrabi-Zadeh, "Diversity maximization via composable coresets," in *CCCG*, 2015, p. 43.