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 the 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 increasing the generalization of an ensemble. Besides, we propose two pruning methods based on diversity management to utilize this relationship, which could increase diversity appropriately and shrink the size of the ensemble without much-decreasing performance. Empirical results validate the reasonableness of the proposed relationship between diversity and ensemble generalization error and the effectiveness of the proposed pruning methods.

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, object tracking [2]–[5], fault diagnosis, malware detection, depression detection [6]–[8], and so on [9]–[19]. 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 [20]–[23]. Dietterich [24] 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 [25]. 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. Existing diversity measures are various, yet none of them show superiority over each other [23], [26], [27]. They are generally divided into pairwise and non-pairwise diversity measures [1], [28]. Lots of previous work made an effort to seek the role of diversity in ensemble learning [29]–[32]. In regression ensembles, the diversity is defined based on the error decomposition [33], in which the error of regression ensembles is split into the accuracy term and diversity term. The classic Ambiguity Decomposition [33] and Bias-Variance-Covariance decomposition [34] are two commonly used error decomposition schemes. However, both of them are only suitable for regression tasks with the square loss.

Moreover, the relationship between diversity and generalization in the literature remains an open question. Some researchers, such as Kuncheva and Whitaker [28], hold doubts about the usefulness of diversity measures in building classification ensembles in real-life pattern recognition problems. Conversely, some researchers hold the view as well that diversity among the members of a team of classifiers is a crucial issue in classifiers’ combination [35]. They are two inconsistent views, so, weirdly, both of them have some supporting experimental results in some cases. For instance, some empirical results [28] showed less correlation between diversity and generalization error by varying the diversity in the ensemble. Meanwhile, some experiments [31] showed precisely the opposite results. Hence, to clarify this issue, whether or not diversity affects the performance of generalization becomes particularly essential.

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 [36], [37]. 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. To extend EPBD to use other diversity measures, we propose an ensemble pruning framework named as “Ensemble Pruning Framework Utilizing the Trade-off Between Accuracy and Diversity (FTAD)” as well, which utilizes the trade-off between accuracy and diversity. Moreover, empirical results are presented to validate the effectiveness of EPBD and FTAD. The contributions in this paper are three-fold:

• A diversity measure based on error decomposition is proposed to quantify the difference among classification ensemble members, for the benefit of conducting theoretical analyses later.
• The relationship between the proposed diversity and the ensemble generalization error is investigated and analyzed theoretically, which demonstrates that diversity has different impacts on the ensemble generalization in different ranges.
• Two ensemble pruning methods are proposed to select a subset of the original ensemble, while one utilizes the proposed theoretical relationship and the other as an extension of the former utilizes the trade-off between accuracy and diversity.

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 and III-E. 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 such as dependence, orthogonality, or complementarity [1], [28]. In this section, we introduce existing methods to generate diverse individual classifiers in ensemble learning. 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. Generating Diverse Individual Classifiers for Ensembles

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 [24], [31]. For example, bagging [38] and Boosting (including many variants) [39]–[41] manipulate input data to promote diversity by choosing different subsets of the training data during training; Random forest [42] manipulates subsets of input data or features to create diversity and gives competitive results; Rotation forest [43] 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, or different learning algorithms. However, the diversity generated by these methods is not precise and could not tell how much diversity in their generating process or whether diversity in ensembles should be increased or not for the specific ensemble.

B. Diversity Measurements in Ensembles

Diversity is usually measured based on the classification results of individual members in an ensemble [26], [30], [35], [44]–[46]. Existing diversity measures are generally divided into pairwise and non-pairwise diversity measures [1], [28]. 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 [23]. Pairwise diversity includes $Q$-statistic [47], $\kappa$-statistic [48], disagreement measure [49], [50], correlation coefficient [51], double-fault [52], and the hamming distance (HD) measure [46]. 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 [23]. It includes interrater agreement [53], Kohavi-Wolpert variance [54], the entropy of the votes [55], [56], the difficulty index [28], [57], the generalized diversity [58], and the coincident failure diversity [58]. 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 [59] in their negative correlation learning (NCL) [60]. 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 [61]. Due to the examined difference or similarity, the ranges of diversity vary between different diversity measures [1]. Moreover, comparing those existing diversity measures is difficult regarding the appropriateness and superiority across them [23]. The performance of a diversity measure might depend on the context of data and the use of diversity [27].

C. The Role that Diversity Plays

Although the crucial role of diversity has been widely accepted, few researchers could tell how diversity works exactly in ensemble methods. In the last decade or so, Brown [29] 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 [30] 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. [31] 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. [32] extended the classic Ambiguity Decomposition from regression problems to classification problems and proved several fundamental properties of the Ambiguity term.
D. Ensemble Pruning

Ensemble pruning, named as ensemble selection or ensemble thinning as well, is proposed to mitigate the shortcoming of ensemble methods that usually require huge memory and processing time regarding the number of individual learners in the ensemble [62]–[68]. It is possible to improve the ensemble generalization performance with a smaller size using ensemble pruning [69], [70]. Selecting the sub-ensembles with the best generalization is quite challenging with exponential computational complexity [35], [71]. Existing ensemble pruning methods could be categorized basically into three general families: ranking-based, clustering-based, and optimization-based. Ranking-based pruning methods sort the learners in the ensemble based on different evaluation function and select the top few of them [72]–[74], with a low searching complexity yet theoretical unsoundness [75]; Clustering-based pruning methods detect groups of learners making similar predictions and prune each group separately [76], [77]; Optimization-based pruning methods utilize different objectives to optimize and find the expected subset with satisfactory generalization performance [78]–[83], requiring an optimization searching algorithm to avoid exhaustive search complexity [75]. Another partition divides ensemble pruning methods into static and dynamic pruning techniques [84]. Static pruning techniques focus on identify a small sub-ensemble with good generalization performance, accompanied by expensive computation [25], [35], [85]–[88]; Dynamic pruning techniques [89]–[91] determine the number of individual classifiers that need to be queried dynamically for each test instance, reducing the time cost yet not reducing the ensemble’s storage requirements [89], [92]–[95]. Besides, Jiang et al. [96] presented a Bayesian framework for ensemble pruning which generated the optimal size of the pruned ensemble first and selected the optimal pruned ensemble accordingly; Krawczyk and Wozniak [97] proposed to do ensemble pruning tasks considering the combination rule training simultaneously.

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, two ensemble pruning methods are 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., \( \mathbf{x} = (x_1, \ldots, x_k) \)) 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 letters (e.g., \( S \)), 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 \( \mathbb{P}, \mathbb{E}, \mathbb{I}, \mathbb{R}, \) and \( \mathbb{Z}^+ \) denote the probability measure, the expectation of a random variable, the indicator function, the real space, and the positive integer space, respectively. The notation \([n]\) represents \( \{1, \ldots, n\} \) for clarity. The notation \((x,y) \in S\) is formally defined by \((x,y) \in S \Leftrightarrow \exists i \in [||S||] : (x_i, y_i) = (x,y)\) for clarity.

A. Measuring Diversity in Classification Ensembles using Error Decomposition

The analysis is built in the PAC (probably approximately correct) framework [98] that one learning task \( D \) corresponds to a probability distribution over the input-output space \( \mathcal{X} \times \mathcal{Y} \). An instance from \( D \) is represented as \((x,y)\), where \( x \in \mathbb{R}^k \) 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 \( F = \{f_1, \ldots, f_{||F||}\} \) denotes the set of these individual classifiers. The predictions of the individual classifiers are combined by weighted voting\(^1\), that is, the output of an ensemble is defined as:

\[
    f_{ens}(x) = \text{sgn} \left( \sum_{i \in [||F||]} w_i \cdot f_i(x) \right),  \tag{1}
\]

where \( w_i \) is the weight coefficient corresponding to the individual classifier \( f_i \), satisfying that \( \sum_{i \in [||F||]} w_i = 1, \forall w_i \geq 0 \). And \( \sum_{i \in [||F||]} w_i f_i(x) = 0 \) indicates a tie in the combination.

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

\[
    \text{margin}(f, x) = f(x) \cdot y.  \tag{2}
\]

As the original error decomposition in regression [33], [99], [100] 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 one single arbitrary instance \( x \) is defined as

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

which is also the discretization of hinge loss function and holds that \( \text{Err}(f, x) = -\frac{1}{2}(f(x)y - 1) \). Therefore, inspired

\(^1\)In the spirit of weighted averaging and plurality voting, each individual classifier votes for a class with the weight, 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. (3) is given zero for binary classification problems.
by the work of [32], [99], 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 $F = \{f_1, \ldots, f_{|F|}\}$ have been trained and are combined by weighted voting $f_{\text{ens}}(x) = \text{sgn} \left( \sum_{i \in [|F|]} w_i f_i(x) \right)$ with $\sum_{i \in [|F|]} w_i = 1$, $\forall w_i \geq 0$. Then for the 0/1 error function, the loss function of the ensemble can be decomposed into

$$
\text{Err}(f_{\text{ens}}, x) = \sum_{i \in [|F|]} w_i \cdot \text{Err}(f_i, x) 
$$

$$
-\frac{1}{2} \left( \text{margin}(f_{\text{ens}}, x) - \sum_{i \in [|F|]} w_i \cdot \text{margin}(f_i, x) \right). 
$$

(4)

Computing the expectation in the data set $S$ yields decomposi-
tion of the generalization error as follows,

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

(5)

where

$$
G = \mathbb{E}_{(x,y) \in S}(\text{Err}(f_{\text{ens}}, x)), 
$$

(6a)

$$
\bar{A} = \sum_{i \in [|F|]} w_i \cdot \mathbb{E}_{(x,y) \in S}(\text{Err}(f_i, x)), 
$$

(6b)

$$
\bar{D} = \frac{1}{2} \mathbb{E}_{(x,y) \in S}(\text{margin}(f_{\text{ens}}, x)) 
$$

$$
-\frac{1}{2} \sum_{i \in [|F|]} w_i \cdot \mathbb{E}_{(x,y) \in S}(\text{margin}(f_i, x)). 
$$

(6c)

**Proof.** For one single arbitrary instance $x$, the error of an ensemble is defined as

$$
\text{Err}(f_{\text{ens}}, x) = -\frac{1}{2} \left( \text{sgn} \left( \sum_{i \in [|F|]} w_i f_i(x) \right) y - 1 \right). 
$$

(7)

Inspired by the error decomposition of regression ensembles, the difference between ensemble error and the average error of individual classifiers is

$$
\text{Err}(f_{\text{ens}}, x) - \sum_{i \in [|F|]} w_i \cdot \text{Err}(f_i, x) 
$$

$$
= \sum_{i \in [|F|]} w_i \cdot \left( \text{Err}(f_{\text{ens}}, x) - \text{Err}(f_i, x) \right) 
$$

$$
= -\frac{1}{2} \sum_{i \in [|F|]} w_i \cdot (f_{\text{ens}}(x) - f_i(x)), 
$$

(8)

and the error decomposition could be rewritten as

$$
\text{Err}(f_{\text{ens}}, x) = \sum_{i \in [|F|]} w_i \cdot \text{Err}(f_i, x) 
$$

$$
-\frac{1}{2} \sum_{i \in [|F|]} w_i (f_{\text{ens}}(x) - f_i(x)) y. 
$$

(9)

The first term, $\sum_{i \in [|F|]} w_i \cdot \text{Err}(f_i, x)$, is the weighted average loss of the individuals. The second term is defined as the diversity term which measures the difference between $f_{\text{ens}}(x)$ and individual classifiers, and it could be rewritten as

$$
\frac{1}{2} \sum_{i \in [|F|]} w_i (f_{\text{ens}}(x) - f_i(x)) y 
$$

$$
= \frac{1}{2} \sum_{i \in [|F|]} w_i (\text{margin}(f_{\text{ens}}, x) - \text{margin}(f_i, x)). 
$$

(10)

Note that this term is corresponding to one specific instance $x$, defined as

$$
\text{div}(f_{\text{ens}}, x) = \frac{1}{2} \text{margin}(f_{\text{ens}}, x) - \frac{1}{2} \sum_{i \in [|F|]} w_i \cdot \text{margin}(f_i, x). 
$$

(11)

And then we obtain the form of the error decomposition for one single instance $x$ as in Eq. (4). Finally, computing the expectation over the data set $S$ will yield the form of the error decomposition for this overall instance set as in Eqs. (5–6). □

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 1** ($\bar{G}, \bar{A}, \bar{D}$ only depend on $\text{div}(f_{ens}, x)$). 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 voting $f_{ens}(x) = \text{sgn}(\sum_{i \in [F]} w_i \cdot f_i(x))$ with $\sum_{i \in [F]} w_i = 1$, $\forall w_i \geq 0$. Then the generalization error $\bar{G}$, accuracy $\bar{A}$ and diversity $\bar{D}$ of this ensemble are all dependent on $\text{div}(f_{ens}, x)$, that is,

\[
\bar{G} = \frac{1}{2} \left(1 - \text{E}_{(x,y) \in S}(\text{sgn}(\lambda - 2 \text{div}(f_{ens}, x)))\right), \\
\bar{A} = \frac{1}{2} \left(1 - \text{E}_{(x,y) \in S}(\lambda - 2 \text{div}(f_{ens}, x))\right), \\
\bar{D} = \text{E}_{(x,y) \in S}(\text{div}(f_{ens}, x)), 
\]

where

\[
\lambda = \begin{cases} 
1, & \text{if } \text{div}(f_{ens}, x) \in (0, \frac{1}{2}) \\
0, & \text{if } \text{div}(f_{ens}, x) = 0 \\
-1, & \text{if } \text{div}(f_{ens}, x) \in (-\frac{1}{2}, 0). 
\end{cases} 
\]

**Proof.** The error of one individual classifier is

\[
\text{Err}(f, x) = -\frac{1}{2} (f(x)y - 1) = -\frac{1}{2} (\text{margin}(f, x) - 1), 
\]

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

\[
\bar{G} = \frac{1}{2} \left(1 - \text{E}_{(x,y) \in S}(\text{margin}(f_{ens}, x))\right), \\
\bar{A} = \frac{1}{2} \left(1 - \text{E}_{(x,y) \in S}(\sum_{i \in [F]} w_i \cdot \text{margin}(f_i, x))\right), \\
\bar{D} = \frac{1}{2} \sum_{i \in [F]} w_i \text{E}_{(x,y) \in S}(\text{margin}(f_{ens}, x) - \text{margin}(f_i, x)). 
\]

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

\[
\text{div}(f_{ens}, x) = \frac{1}{2} \text{sgn}\left(\sum_{i \in [F]} w_i f_i(x)\right) y - \frac{1}{2} \sum_{i \in [F]} w_i \text{margin}(f_i, x) \\
= \frac{1}{2} \text{sgn}\left(\sum_{i \in [F]} w_i \text{margin}(f_i, x)\right) - \frac{1}{2} \sum_{i \in [F]} w_i \text{margin}(f_i, x), 
\]

then define that

\[
\text{margin}(f_{ens}, x) \overset{\text{def}}{=} \sum_{i \in [F]} w_i \cdot \text{margin}(f_i, x), 
\]

and consequently obtain that for clarity,

\[
\text{margin}(f_{ens}, x) = \lambda - 2 \text{div}(f_{ens}, x), 
\]

where $\lambda$ is defined in Eq. (13). Finally, we could obtain the Eqs. (12a)-(12c) as described in Corollary 1.

\[\Box\]

C. Relationship Between Diversity and Ensemble Performance

In this subsection, we intend to explain whether and how diversity affects the generalization performance of classification ensembles. Herbrich and Graepel [36], [37] 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 $(x, y) \sim (X, Y) \in (\mathcal{X}, \mathcal{Y})$ in $S$ is drawn independent and identically distributed (i.i.d.) according to a certain unknown probability measure $P_{Y|X}P_X$, and we only consider linear classifiers as ensembles just like $f_{ens}(\cdot)$ in Eq. (1), i.e.,

\[
\mathcal{F} = \{x \mapsto \text{sgn}(\langle w, \phi(x) \rangle) \mid w \in \mathcal{W}\}, 
\]

where $\phi(x) = [f_1(x), ..., f_F(x)]^T$ and $w = [w_1, ..., w_F]^T$. Notice that $\sum_{i \in [F]} w_i = 1$ leads to a one-to-one correspondence of hypotheses $f_{ens}(x, w) \in \mathcal{F}$ to their parameters $w \in \mathcal{W}$. Assume that the existence of a “true” hypothesis $w^* \in \mathcal{W}$ labeling the data, leading to a PAC-likelihood,

\[
P_{Y|X=x}(y) = \mathbb{I}(y = \text{sgn}(\langle w^*, \phi(x) \rangle)). 
\]

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

\[
V(S) = \{w \in \mathcal{W} \mid \forall (x, y) \in S, \text{sgn}(\langle w, \phi(x) \rangle) = y\}. 
\]

The true risk $R(w)$ of consistent hypotheses $w \in V(S)$ is

\[
R(w) = \mathbb{E}_{X,Y}[\text{I}(\text{sgn}(\langle w, \phi(X) \rangle) \neq Y)], 
\]

and a concept of the margin $\gamma(w)$ of an ensemble $w$ is introduced [36], [37] as

\[
\gamma(w) = \min_{(x,y) \in S} y(w, \phi(x)), 
\]

since all ensemble classifiers, $w \in V(S)$ are indistinguishable in terms of error rate on the given training set $S$. Herbrich and Graepel [37] presented that the generalization error $R(w)$ of an ensemble was bounded from above in terms of the margin $\gamma(w)$, as shown in Lemma 1.

**Lemma 1** (Relationship between margin and generalization [37]). For any probability measures $P_X$ such as $P_X(\|\phi(x)\| \leq \delta) = 1$ and for any $\xi \in (0, 1]$, with probability at least $1 - \xi$ over the random draw of the training instances $S \in (X, Y)^{|S|}$, for any arbitrary consistent ensemble classifier $w \in V(S)$ with a positive margin $\gamma(w) > \sqrt{32\delta^2/|S|}$, the generalization error $R(w)$ could be bounded from above by

\[
R(w) \leq \frac{2}{|S|} \left(\kappa(w) \log_2 \left(\frac{8e|S|}{\kappa(w)}\right) \log_2 (32|S|) + \log_2 \left(\frac{2|S|}{\xi}\right)\right), 
\]

where $\kappa(w) = \left(\frac{8\delta}{\gamma(w)}\right)^2$.

Subsequently, inspired by the relationship between margin $\gamma(w)$ and generalization error $R(w)$ in Lemma 1 [37], we present the following relationship between the proposed diversity and generalization. It is worth mentioning that the margin $\gamma(w)$ could be linked with the proposed diversity $\text{div}(f_{ens}, x)$
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., \( \kappa(w) \log_2 \left( \frac{8e|S|}{\kappa(w)} \right) \), described as \( \hat{R}(w) \), that is,

\[
\hat{R}(w) \overset{\text{def}}{=} \kappa(w) \log_2 \left( \frac{8e|S|}{\kappa(w)} \right) \approx \left( \frac{1}{8\delta} \right) \left( 1 - 2\varepsilon \right) \left( \lambda - 2 \text{div}(f_{ens}, x^*) \right)^2 \cdot \log_2 \left( \frac{8e|S|}{\left( 1 - 2\varepsilon \right) \left( \lambda - 2 \text{div}(f_{ens}, x^*) \right)^2} \right),
\]

(30)

with the same variation tendency as \( R(w) \).

According to Theorem 2, the generalization error of the ensemble could be quantified by the diversity. Eq. (30) reflects the relationship of individual classifiers \( \text{div}(f_{ens}, x^*) \in (-\frac{1}{2}, \frac{1}{2}) \) and the performance of generalization error of the ensemble \( \hat{R}(w) \). Therefore, the first and the second derivative\(^2\) of the estimated risk \( \hat{R}(w) \) could be obtained by taking the derivative of \( \hat{R}(w) \) on \( \text{div}(f_{ens}, x^*) \), i.e.,

\[
\hat{R}' = 4 \left( \frac{1}{8\delta} \right) \left( 1 - 2\varepsilon \right) \left( \lambda - 2 \text{div}(f_{ens}, x^*) \right)^3 \cdot \log_2 \left( \frac{8e|S|}{\left( 1 - 2\varepsilon \right) \left( \lambda - 2 \text{div}(f_{ens}, x^*) \right)^2} \right),
\]

(31)

\[
\hat{R}'' = 8 \ln 2 \left( \frac{1}{8\delta} \right) \left( 1 - 2\varepsilon \right) \left( \lambda - 2 \text{div}(f_{ens}, x^*) \right)^4 \cdot \left( 3 \ln \left( \frac{8e|S|}{\left( 1 - 2\varepsilon \right) \left( \lambda - 2 \text{div}(f_{ens}, x^*) \right)^2} \right) - 2 \right).
\]

(32)

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

---

2The 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 that function's derivative, 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 [101].
Then the stagnation point of the first derivative
\( q_2 \) is found by making \( \frac{dR^1}{dw}(w) = 0 \). 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 \( q_4 \) is found by making \( \frac{dR^2}{dw}(w) = 0 \), to analyze the concavity of the function \( R(w) \); the stagnation point of the second derivative \( q_6 \) is found by making \( \frac{dR^3}{dw}(w) = 0 \), to analyze the concavity of the function \( R(w) \). Their values are

\[
\begin{align*}
q_1 &= \varepsilon, \\
q_2 &= \frac{1}{2} \left( 1 - \frac{\varepsilon}{1 - 2\varepsilon} \right), \\
q_4 &= \frac{1}{2} \left( 1 - \frac{\delta}{1 - 2\varepsilon} \right) \left( \frac{8}{|S|} \right), \\
q_5 &= \frac{1}{2} \left( 1 - \frac{\delta}{1 - 2\varepsilon} \right) \left( \frac{8}{|S|} \right), \\
q_6 &= \frac{1}{2} \left( 1 - \frac{\delta}{1 - 2\varepsilon} \right) \left( \frac{8}{|S|} \right).
\end{align*}
\]

respectively, where an implied condition exists, i.e.,

\[
\varepsilon \leq \frac{\delta}{1 - 2\varepsilon} \leq 1 - 2\varepsilon.
\]

Note that \( q_1 < q_2 < q_3 \) and \( q_6 < q_5 < q_4 = q_2 \). Then the domain of diversity \( \text{div}(f_{ens}, x^*) \) could be divided into eight intervals, as shown in Table I. Up to now, we can analyze in detail the effect of diversity on generalization error within different intervals.

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

(2) In real situations, \( \text{div}(f_{ens}, x^*) \) is negative in most cases which means that “the ensemble classifier has misclassified the instance that corresponds to the selected \( \text{div}(f_{ens}, x^*) \).” Notice that in the finite experiments, when the value of \(|S|\) satisfies Eq. (34), it is rare that \( \text{div}(f_{ens}, x^*) \) is larger than zero. If \( \text{div}(f_{ens}, x^*) < 0 \), increasing diversity will improve the ensemble’s performance, which is the reason why we should increase diversity in that case.

(3) When \( \text{div}(f_{ens}, x^*) \) 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 \( \text{div}(f_{ens}, x^*) \) is to zero, the greater their advantage is. In the range near 1/2, individual classifiers that classify the instance correctly take little advantage, and the closer \( \text{div}(f_{ens}, x^*) \) is to 1/2, the less 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 \( \text{div}(f_{ens}, x^*) \) 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 falls 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 \( \text{div}(f_{ens}, x^*) \) to get a more effective ensemble classifier. However, in other ranges, it is not worth increasing diversity, and the ensemble could remain the same to continue the learning process.

### D. Ensemble Pruning based on Diversity and One Ensemble Pruning Framework

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.

**Algorithm 1 Ensemble Pruning based on Diversity (EPBD)**

**Input:** A training set \( S = \{(x_i, y_i) \} \) \( \in |S| \), an original ensemble \( F = \{f_j \} \) \( \in |F| \), and the threshold \( \alpha \).

**Output:** The pruned sub-ensemble \( H \) \( (H \subset F) \).

1. \( H = \emptyset \).
2. **repeat**
3. **search** for the specific data instance \( (x, y) \) which satisfies the search criterion Eq. (13).
4. **sort** classifiers in \( F \) that classify this instance correctly in ascending order according to the accuracy performance.
5. Move the top one \( f(\cdot) \) in the previous step from \( F \) to \( H \).
6. **until** The termination condition is satisfied.

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), and (iii) the ratio \( \alpha \) which means the individual classifiers in the pruned sub-ensemble are chosen in a ratio of \( \alpha(\%) \) from the original ensemble. Notice that all individual members are treated equally here, which means that \( \bar{w}_i = 1/|F|, \forall i \in |F| \). For example, individual members in one ensemble generated by AdaBoost might have different values of
weights in the combination, however, their weight coefficients would still be treated as important as each other in the pruning process, to simplify the analyses since it doesn’t talk more about the situation where the individual classifiers in the ensemble are not treated equally. 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., $|F|$), 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).

The objective of this method is to obtain a pruned sub-ensemble that approaches the performance as optimal as possible. As shown in Algorithm 1, it firstly picks out an individual classifier considering diversity and accuracy simultaneously, raising diversity in line 3 and increasing accuracy in line 4. Specifically, the sorting principle we use in line 4 is mainly based on the accuracy performance of individual classifiers. 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 expected classifiers in $F$ repeatedly until the termination conditions are met.

Moreover, inspired by the proposed ensemble pruning method (EPBD), we propose an ensemble pruning framework by utilizing the trade-off between accuracy and diversity, named as “Ensemble Pruning Framework Utilizing the Trade-off Between Accuracy and Diversity (FTAD),” presented in Algorithm 2. The diversity measure used in FTAD could be any existing diversity measure, such as $Q$-statistic or the disagreement measure. Note that EPBD is one particular case of FTAD by using the proposed diversity measure in this paper.
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:

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

2. Secondly, the complexity of line 4 is $O(j \log j)$ where $j = n - i + 1$, when $i \in \{1, \ldots, k\}$.

3. Thirdly, the complexity of line 5 is $O(1)$.

Therefore, the overall computational complexity of EPBD is $O(\sum_{i=0}^{h} (jm + m \log m + j \log j))$, i.e., $O(km(n-k) + km \log m + \sum_{i=0}^{n} (i \log i))$.

IV. EXPERIMENTS

In this section, we elaborate on our experiments to evaluate the proposed relationship between the proposed diversity and ensemble generalization performance and the proposed algorithms. The data sets we use include one image data set with 12,500 pictures (Dogs vs Cats\(^3\)) and 28 data sets from UCI repository\(^4\) [103]. 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 [38] and AdaBoost [41], [104] 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, linear SVMs (LSVM), support vector machines (SVM), and multilayer perceptrons (MLP). 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) [69], orientation ordering pruning (OO) [105], diversity regularized ensemble pruning (DREP) [35], accuracy-based ensemble pruning (TSP.AP) [73], and diversity-based ensemble pruning (TSP.DP) [73] as well as optimization-based methods, namely single-objective ensemble pruning (SEP), and Pareto ensemble pruning (PEP) [81], the maximal relevant minimal redundant (MRMR) method [82], the maximal relevancy minimum complementary (MRMC) method [82], and the discriminative ensemble pruning (DiscEP) [83]. Moreover, we adopt two methods from diversity maximization via composable coresets [106], [107] 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 evaluated on the test set. It is worth mentioning that several methods (such as OO, DREP, SEP, PEP, LCS, and TSP.AP) cannot fix the number of classifiers after ensemble pruning, while others could do that by giving a pruning rate that is the upper 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 increasing or reducing the size of pruned sub-ensembles and affect their space cost.

A. Validating the Error Decomposition for Classification Ensembles

In this subsection, experiments are conducted to verify the proposed error decomposition for classification ensembles (Theorem 1). To verify the effectiveness of the proposed diversity measure in Eq. (6), we compare the proposed diversity measure with other diversity measures in the literature, including the disagreement measure (Disag) [49], [50], \(Q\)-statistic (QStat) [47], correlation coefficient (Corre) [51], \(\kappa\)-statistic (KStat) [48], double-fault (DoubF) [52], Kohavi-Wolpert variance (KWVar) [54], the interrater agreement (Inter) [53], the entropy of the votes (EntCC [55] and EntSK [56]), the difficulty index (Diffi) [28], [57], the generalized diversity (GeneD) [58], and the coincident failure diversity (CFail) [58]. As shown in

---

\(^3\)http://www.kaggle.com/c/dogs-vs-cats

\(^4\)http://archive.ics.uci.edu/ml/datasets.html
Table II

Comparison of the state-of-the-art pruning methods with EPBD using bagging to produce homogeneous ensembles. The column named "ENSEM" is the corresponding result for the entire ensemble without pruning. The best accuracy with a lower standard deviation is indicated with bold fonts for each data set (ROW).

| Data Set          | Ensemble | ES | KL | KP | OO | DREP | SEP | PEP | MRMR | MRMG | DiscEP | TSP.AP | GMA | LCS | EPBD |
|-------------------|----------|----|----|----|----|------|-----|-----|------|------|-------|-------|-----|-----|------|
| waveform          | 98.15     | 90.01 | 91.27 | 86.34 | 84.67 | 86.30 | 84.20 | 91.20 | 88.61 | 90.95 | 90.10 | 88.61 | 90.24 | 88.61 | 90.10 | 88.61 |
| heart             | 55.56     | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 |
| iono              | 94.02     | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 |
| shuttle           | 96.75     | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 |
| page              | 90.40     | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 |

**Table III**

Comparison of the state-of-the-art pruning methods with EPBD using AdaBoost to produce homogeneous ensembles. The column named "ENSEM" is the corresponding result for the entire ensemble without pruning. The best accuracy with a lower standard deviation is indicated with bold fonts for each data set (ROW).

| Data Set     | Ensemble | ES | KL | KP | OO | DREP | SEP | PEP | MRMG | MRMG | DiscEP | TSP.AP | GMA | LCS | EPBD |
|--------------|----------|----|----|----|----|------|-----|-----|------|------|-------|-------|-----|-----|------|
| iono         | 94.02     | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 |
| waveform      | 98.15     | 90.01 | 91.27 | 86.34 | 84.67 | 86.30 | 84.20 | 91.20 | 88.61 | 90.95 | 90.10 | 88.61 | 90.24 | 88.61 | 90.10 | 88.61 |
| heart         | 55.56     | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 | 55.56 |
| iono          | 94.02     | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 | 94.02 |
| shuttle       | 96.75     | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 | 96.75 |
| page          | 90.40     | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 | 90.40 |

5 The correlation coefficient used in this paper refers to the Pearson product-moment correlation coefficient. It measures the linear dependence between two variables X and Y. Its value ranges from -1 to 1. A value of 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation [109].

Figure 5(a) and Figure 3, the degree of correlation between the loss difference \((G - A)\) and the proposed diversity measure \(D\) in Eq. (6c) is higher than that of the loss difference and other diversity measures. Similar results are presented in Figure 5(b) and Figure 4. Therefore, we could suppose the proposed error decomposition for classification ensembles is reasonable.

B. Validating the Proposed Relationship between Diverse Performance and Ensemble Performance

In this subsection, experiments are conducted to verify the proposed relationship between the proposed diversity and generalization performance of classification ensembles (Theorem 2). Before that, we need to verify the ensemble performance’s relationship with the estimated risk. The experimental results are reported in Figure 6, including different individual classifiers. In each experiment, the 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 6(a)–(b), presenting a high level of correlation between them. Besides, experimental results reported in Figures 7–8 indicate that the relationship between the proposed diversity and ensemble performance in Theorem 2 is faithful. Along with the increase of diversity, the ensemble performance would show an “upward-downward-upward trend” in Figures 7(a) and 8(a), which coincides with the analyses in Section III-C. Therefore, we 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.

C. 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, MRMG, MRMG, DiscEP, TSP.AP, TSP.DP, GMA, and LCS with the proposed EPBD method. Experimental results reported in Tables II–III contain the average test accuracy of each method and the corresponding standard deviation under 5-fold cross-validation on each data set. For instance, each row (data set) in Table II compares the classification accuracy using bagging with the same type of individual classifiers, indicating results include higher accuracy and lower standard deviation by bold fonts. Besides, the significance of the difference in the accuracy performance between the two methods is examined by two-tailed paired t-tests 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, the one with higher values of accuracy will win. The performance of each method is reported in the last two rows of Table II, compared with EPBD in terms of the average rank and the number of data sets that EPBD has won, tied, or lost, respectively. We may notice that OO or PEP might achieve the best results instead of EPBD in a few cases. Yet EPBD could still reach an equal level of accuracy performance without significant difference ending up with a tie. Moreover, both of them cannot restrict the number of classifiers after ensemble pruning like EPBD does, leading to keeping more individual classifiers than we desire in most situations. 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 effective and that utilizing our proposed relationship is reasonable. Figure 9(a) shows that EPBD could achieve competitive results as OO, PEP, MRMR, and TSPAP, and EPBD has significant superiority over other compared pruning methods. Figure 9(b) presents the aggregated rank for each method, describing the similar conclusions to Figure 9(a). Figures 9(c)–9(d) illustrate the aggregated time cost and the aggregated space cost for each pruning method, respectively. It could be inferred that PEP, DiscEP, and LCS usually cost huge processing time to do ensemble pruning tasks. Similar results are reported in Figure 10.

D. Comparison Between EPBD and FTAD Using Other Diversity Measures

In this subsection, we compare the performance of FTAD with various diversity measures included, involving pairwise measures (Disag, QStat, Corre, KStat, and DoubF) and non-pairwise measures (KVar, Inter, EntCC, EntSK, Difi, GeneD, and CFail) with the proposed EPBD method. Experimental results reported in Tables IV–VI contain the average test performance of each method.
Experimental results are reported in Figure 13. As shown in Figure 13, 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. More comparison of space cost could be referred to Figure 9(d) and Figure 10(d). 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 pairwise diversity and generalization. These theoretical analyses solve an open question in the ensemble learning community. First of all, inspired by the regression ensembles, this paper proposed the measure of pairwise diversity and generalization. These theoretical analyses solve an open question in the ensemble learning community.
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 \textit{EPBD} that can prune an ensemble without much performance degradation. Besides, an ensemble pruning framework (i.e., \textit{FTAD}) is proposed to utilize the trade-off between accuracy and diversity and generate pruned sub-ensembles with a smaller size. Note that in \textit{FTAD}, not only the proposed diversity measure in this paper but also other existing diversity measures could be employed. Although this generalization bound could be loose, our work provides a direction and one theoretical attempt to reveal the impacts of diversity on the generalization. Our future work is to generalize the proposed methodology to the multi-classification problems.

**APPENDIX**

**EXAMPLE OF EXTENDED ANALYSES IN MULTI-CLASSIFICATION PROBLEMS**

In this section, we give a short example to extend the aforementioned analyses in Section III to multi-classification problems, that means $\mathcal{Y} = \{0, 1, \ldots, n_c - 1\}$ with $n_c \geq 2$, $n_c \in \mathbb{Z}^+$ different labels. Individual classifiers in an ensemble
\((F = \{f_1, \ldots, f_{|F|}\})\) have been trained and are combined by weighted voting, i.e.,

\[
\text{f}_{\text{ens}}(x) = \arg \max_{i \in \mathcal{Y}} \sum_{i \in [F]} w_i \cdot c(x_i) = c_k ,
\]

(35)

where \(w_i\) is the weight corresponding to the individual classifier \(f_i\), satisfying that \(\sum_{i \in [F]} w_i = 1, \forall w_i \geq 0\). To extend the aforementioned analyses in Section III to multi-classification problems, we give the extended proposed error decomposition and diversity measure as a short example. To this end, we need to extend the \textit{margin} of an ensemble to multi-classification problems first. Inspired by [110], we introduce a new concept named as \textit{"credence"} of one classifier \(f(\cdot)\), represented by \textit{cred(\cdot)}, as mappings from \(\mathcal{X} \times \mathcal{Y}\) to \([0, 1]\), with the interpretation that the result \(f(x)\) predicted by \(f\) is more plausible for the label \(y\) of the instance \(x\) if cred(\(f, x, y\)) is closer to 1. Besides, the \textit{credence} of the ensemble \(f_{\text{ens}}\) on one instance \((x, y)\) is defined as

\[
\text{cred}(f_{\text{ens}}, x, y) = \sum_{i \in [F]} w_i \cdot \text{cred}(f_i, x, y) ,
\]

(36)

therefore the ensemble predicts label \(y\) for the instance \(x\) if \(\text{cred}(f_{\text{ens}}, x, y) > \max_{y' \neq y, y' \in \mathcal{Y}} \text{cred}(f_{\text{ens}}, x, y')\) (and ties are broken arbitrarily). Subsequently, the \textit{margin} of the ensemble on the instance is defined as

\[
\text{margin}(f_{\text{ens}}, x) = \text{cred}(f_{\text{ens}}, x, y) - \max_{y' \neq y, y' \in \mathcal{Y}} \text{cred}(f_{\text{ens}}, x, y') ,
\]

(37)

and that of one individual classifier \(f_i\) could be obtained similarly. Consequently, \(f_{\text{ens}}\) would give the wrong prediction on \((x, y)\) only if \(\text{margin}(f_{\text{ens}}, x) \leq 0\). Similar to Section III-A, we define the error function as

\[
\text{Err}(f_{\text{ens}}, x) = \frac{1}{2} (1 - \text{margin}(f_{\text{ens}}, x)) ,
\]

(38)

and the diversity measure would be

\[
\text{div}(f_{\text{ens}}, x) = \sum_{i \in [F]} w_i \cdot \text{Err}(f_i, x) - \text{Err}(f_{\text{ens}}, x) .
\]

(39)

The remaining issue is the definitions of \textit{credence}. As we know in binary classification problems where \(\mathcal{Y} = \{-1, +1\}\), the \textit{credence} of one classifier \(f\) is supposed to be viewed as

\[
\text{cred}(f, x, y) = \frac{f(x) + 1}{2} .
\]

(40)

As for multi-classification problems where \(\mathcal{Y} = \{0, 1, \ldots, n_c - 1\}\) \((n_c \geq 2, n_c \in \mathbb{Z}^+)\), three functions could be proposed as possible forms of \textit{credence}, i.e.,

\[
\begin{align*}
\text{cred}_1(f, x, y) &= 1 \cdot I(f(x) = y) , \\
\text{cred}_2(f, x, y) &= 1 - \frac{|f(x) - y|}{n_c - 1} , \\
\text{cred}_3(f, x, y) &= 1 - \frac{(f(x) - y)^2}{(n_c - 1)^2} .
\end{align*}
\]

(41a-b-c)

Consequently, the difference between the ensemble and its individual members, defined as the “diversity” measure, could be obtained by

\[
\text{div}(f_{\text{ens}}, x) = - \left( \text{Err}(f_{\text{ens}}, x) - \sum_{i \in [F]} w_i \cdot \text{Err}(f_i, x) \right)
\]

\[
= \sum_{i \in [F]} w_i \cdot \text{margin}(f_{\text{ens}}, x) - \text{margin}(f_i, x)
\]

\[
= \sum_{i \in [F]} w_i \cdot \frac{\text{cred}(f_{\text{ens}}, x, y) - \text{cred}(f_i, x, y)}{2}
\]

\[
- \sum_{i \in [F]} \frac{w_i}{2} \left( \max_{y' \neq y} \text{cred}(f_{\text{ens}}, x, y') - \max \text{cred}(f_i, x, y') \right) ,
\]

(42)

that is,

\[
\text{div}(f_{\text{ens}}, x) = \frac{1}{2} \sum_{i \in [F]} w_i \cdot \max_{y' \neq y} \text{cred}(f_i, x, y')
\]

\[
- \frac{1}{2} \max \text{cred}(f_{\text{ens}}, x, y') .
\]

(43)

Analogous results like Theorem 1 could be obtained subsequently. Notice that the appendix is to point out one possibility of extending our theoretical results to multi-classification problems, which is a start. Further investigations and careful experiments are still needed before getting a more promising conclusion.

\textbf{References}

[1] Z.-H. Zhou, \textit{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 \textit{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 \textit{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 \textit{ICIP}, Oct 2014, pp. 843–847.

[5] H. Ykhlef and D. Bouchaffra, “An efficient ensemble pruning approach based on simple coalition games,” \textit{Inform Fusion}, vol. 34, pp. 28–42, 2017.

[6] J. Sun, L. Li, and Q. Hu, “Fault diagnosis based on pruned ensemble,” in \textit{ICMLC}, vol. 1, 2012, pp. 35–40.

[7] J. H. Abawajy, M. Chowdhury, and A. Kelarev, “Hybrid consensus pruning of ensemble classifiers for big data malware detection,” \textit{IEEE T Cloud Comput}, vol. 8, no. 2, pp. 398–407, 2020.

[8] Z. Liu, C. Li, X. Gao, G. Wang, and J. Yang, “Ensemble-based depression detection in speech,” in \textit{IEEE BBM}, 2017, pp. 975–980.

[9] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen, “Feature selection inspired classifier ensemble reduction,” \textit{IEEE T Cybernetics}, vol. 44, no. 8, pp. 1259–1268, 2014.

[10] Z. Sun, Q. Song, and X. Zhu, “Using coding-based ensemble learning to improve software defect prediction,” \textit{IEEE T Syst Man Cy C}, vol. 42, no. 6, pp. 1806–1817, 2012.

[11] W.-Y. Lin, Y.-H. Hu, and C.-F. Tsai, “Machine learning in financial crisis prediction: A survey,” \textit{IEEE T Syst Man Cy C}, vol. 42, no. 4, pp. 421–436, 2012.

[12] A. Dutta and P. Dasgupta, “Ensemble learning with weak classifiers for fast and reliable unknown terrain classification using mobile robots,” \textit{IEEE T Syst Man Cy S}, vol. 47, no. 11, pp. 2933–2944, 2017.

[13] D. Guo, Y. Jin, J. Ding, and T. Chai, “Heterogeneous ensemble-based infill criterion for evolutionary multiobjective optimization of expensive problems,” \textit{IEEE T Cybernetics}, vol. 49, no. 3, pp. 1012–1025, 2019.

[14] Y. Bi, B. Xue, and M. Zhang, “Genetic programming with a new representation to automatically learn features and evolve ensembles for image classification,” \textit{IEEE T Cybernetics}, pp. 1–15, 2020.

[15] J. Zhuang, K. Geng, and G. Yin, “Ensemble learning based brain-computer interface system for ground vehicle control,” \textit{IEEE T Syst Man Cy S}, pp. 1–13, 2019.
