Exploiting Unlabeled Data to Enhance Ensemble Diversity

Min-Ling Zhang[*,†]
[*]*School of Computer Science and Engineering, Southeast University, Nanjing 210096, China*
Email: zhangml@seu.edu.cn

Zhi-Hua Zhou[†]
[†]*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China*
Email: zhouzh@lamda.nju.edu.cn

***Abstract*—Ensemble learning aims to improve generalization ability by using multiple base learners. It is well-known that to construct a good ensemble, the base learners should be *accurate* as well as *diverse*. In this paper, unlabeled data is exploited to facilitate ensemble learning by helping augment the diversity among the base learners. Specifically, a semi-supervised ensemble method named UDEED is proposed. Unlike existing semi-supervised ensemble methods where error-prone *pseudo-labels* are estimated for unlabeled data to enlarge the labeled data to improve accuracy, UDEED works by maximizing accuracies of base learners on labeled data while maximizing diversity among them on unlabeled data. Experiments show that UDEED can effectively utilize unlabeled data for ensemble learning and is highly competitive to well-established semi-supervised ensemble methods.**

***Keywords*-ensemble learning; unlabeled data; diversity**

I. INTRODUCTION

In *ensemble learning* [8], a number of base learners are trained and then combined for prediction to achieve strong generalization ability. Numerous effective ensemble methods have been proposed, such as BOOSTING [9], BAGGING [4], STACKING [19], etc., and most of these methods work under the supervised setting where the labels of training examples are known. In many real-world tasks, however, unlabeled training examples are readily available while obtaining their labels would be fairly expensive. *Semi-supervised learning* [5] is a major paradigm to exploit unlabeled data together with labeled training data to improve learning performance automatically, without human intervention.

This paper deals with semi-supervised ensembles, that is, ensemble learning with labeled and unlabeled data. In contrast to the huge volume of literatures on ensemble learning and on semi-supervised learning, only a few work has been devoted to the study of semi-supervised ensembles. As indicated by Zhou [20], this was caused by the different philosophies of the ensemble learning community and the semi-supervised learning community. The ensemble learning community believes that it is able to boost the performance of weak learners to strong learners by using multiple learners, and so there is no need to use unlabeled data; while the semi-supervised learning community believes that it is able to boost the performance of weak learners to strong learners by exploiting unlabeled data, and so there is no need to use multiple learners. However, as Zhou indicated [20], there are several important reasons why ensemble learning and semi-supervised learning are actually mutually beneficial, among which an important one is that by considering unlabeled data it is possible to help augment the *diversity* among the base learners, as explained in the following paragraph.

It is well-known that the generalization error of an ensemble is related to the average generalization error of the base learners and the diversity among the base learners. Generally, the lower the average generalization error (or, the higher the average accuracy) of the base learners and the higher the diversity among the base learners, the better the ensemble [11]. Previous ensemble methods work under supervised setting, trying to achieve a high average accuracy and a high diversity by using the labeled training set. It is noteworthy, however, pursuing a high accuracy and a high diversity may suffer from a dilemma. For example, for two classifiers which have perfect performance on the labeled training set, they would not have diversity since there is no difference between their predictions on the training examples. Thus, to increase the diversity needs to sacrifice the accuracy of one classifier. However, when we have unlabeled data, we might find that these two classifiers actually make different predictions on unlabeled data. This would be important for ensemble design. For example, given two pairs of classifiers, (A, B) and (C, D), if we know that all of them are with 100% accuracy on labeled training data, then there will be no difference taking either the ensemble consisting of (A, B) or the ensemble consisting of (C, D); however, if we find that A and B make the same predictions on unlabeled data, while C and D make different predictions on some unlabeled data, then we will know that the ensemble consisting of (C, D) should be better. So, in contrast to previous ensemble methods which focus on achieving both high accuracy and high diversity using only the labeled data, the use of unlabeled data would open a promising direction for designing new ensemble methods.

In this paper, we propose the UDEED (*Unlabeled Data to Enhance Ensemble Diversity*) approach. Experiments show that by using unlabeled data for diversity augmentation, UDEED achieves much better performance than its counterpart which does not consider the usefulness of unlabeled data. Moreover, UDEED also achieves highly comparable performance to other state-of-the-art semi-supervised ensem-

ble methods.

The rest of this paper is organized as follows. Section II briefly reviews related work on semi-supervised ensembles. Section III presents UDEED. Section IV reports our experimental results. Finally, Section V concludes.

II. Related Work

As mentioned before, in contrast to the huge volume of literatures on ensemble learning and on semi-supervised learning, only a few work has been devoted to the study of semi-supervised ensembles.

Zhou and Li [21] proposed the TRI-TRAINING approach which uses three classifiers and in each round if two classifiers agree on an unlabeled instance while the third classifier disagrees, then the two classifiers, under a certain condition, will label this unlabeled instance for the third classifier; the three classifiers are voted to make prediction. This is a *disagreement-based* semi-supervised learning approach [22], which can be viewed as a variant of the famous *co-training* method [3]. Later, Li and Zhou [14] extended TRI-TRAINING to CO-FOREST, by including more base classifiers and in each round the *majority teach minority* strategy is still adopted.

In addition to TRI-TRAINING and CO-FOREST, there are several *semi-supervised boosting* methods [1], [6], [7], [16], [18]. D'Alché Buc et al. [7] proposed SSMBOOST to handle unlabeled data within the margin cost functional optimization framework for boosting [17], where the margin of an ensemble H on unlabeled data $\boldsymbol{x}$ is defined as either $H(\boldsymbol{x})^2$ or $|H(\boldsymbol{x})|$. Furthermore, SSMBOOST requires the base learners to be semi-supervised algorithms themselves. Later, Bennett et al. [1] developed ASSEMBLE, which labels unlabeled data $\boldsymbol{x}$ by the current ensemble as $y = \text{sign}\,[H(\boldsymbol{x})]$, and then iteratively puts the newly labeled examples into the original labeled set to train a new base classifier which is then added to H. Following the same margin cost functional optimization framework, Chen and Wang [6] added a local smoothness regularizer to the objective function used by ASSEMBLE to help induce new base classifier with a more reliable self-labeling process. Other than the margin cost functional formalization, MCSSB [18] and SEMIBOOST [16] estimate the labels of unlabeled instances by optimizing an objective function containing two terms. The first term encodes the *manifold assumption* that unlabeled instances with high similarities in input space should share similar labels, while the other term encodes the *clustering assumption* that unlabeled instances with high similarities to a labeled example should share its given label. The difference lies in that MCSSB [18] implemented the objective terms based on Bregman divergence while SEMIBOOST [16] implemented them with traditional exponential loss.

A commonness of these existing semi-supervised ensemble methods is that they construct ensembles iteratively, and in particular, the unlabeled data are exploited through assigning *pseudo-labels* for them to enlarge labeled training set. Specifically, pseudo-labels of unlabeled instances are estimated based on the ensemble trained so far [1], [7], [14], [21], or with specific form of smoothness or manifold regularization [6], [16], [18]. After that, by regarding the estimated labels as their *ground-truth* labels, unlabeled instances are used in conjunction with labeled examples to update the current ensemble iteratively.

Although various strategies have been employed to make the pseudo-labeling process more reliable, such as by incorporating data editing [13], the estimated pseudo-labels may still be prone to error, especially in initial training iterations where the ensemble is only moderately accurate. In the next section we will present the UDEED approach. Rather than working with pseudo-labels to enlarge labeled training set, UDEED utilizes unlabeled data in a different way, i.e., help augment the *diversity* among base learners.

III. The UDEED Approach

A. General Formulation

Let $\mathcal{X} = \mathcal{R}^d$ be the d-dimensional input space and $\mathcal{Y} = \{-1, +1\}$ be the output space. Suppose $\mathcal{L} = \{(\boldsymbol{x}_i, y_i) | 1 \leq i \leq L\}$ contains L *labeled* training examples and $\mathcal{U} = \{\boldsymbol{x}_i | L+1 \leq i \leq L+U\}$ contains U *unlabeled* training examples, where $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. In addition, we use $\tilde{\mathcal{L}} = \{\boldsymbol{x}_i | 1 \leq i \leq L\}$ to denote the unlabeled data set derived from $\mathcal{L}$.

We assume that the classifier ensemble is composed of m base classifiers $\{f_k | 1 \leq k \leq m\}$, where each of them takes the form $f_k : \mathcal{X} \to [-1, +1]$. Here, the real value of $f_k(\boldsymbol{x})$ corresponds to the confidence of $\boldsymbol{x}$ being positive. Accordingly, $(f_k(\boldsymbol{x})+1)/2$ can be regarded as the *posteriori probability* of being positive given $\boldsymbol{x}$, i.e. $P(y = +1|\boldsymbol{x})$.

The basic idea of UDEED is to maximize the fit of the classifiers on the labeled data, while maximizing the diversity of the classifiers on the unlabeled data. Therefore, UDEED generates the classifier ensemble $\boldsymbol{f} = (f_1, f_2, \cdots, f_m)$ by minimizing the following loss function:

$$V(\boldsymbol{f}, \mathcal{L}, \mathcal{D}) = V_{emp}(\boldsymbol{f}, \mathcal{L}) + \gamma \cdot V_{div}(\boldsymbol{f}, \mathcal{D}) \qquad (1)$$

Here, the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ corresponds to the *empirical loss* of $\boldsymbol{f}$ on the labeled data set $\mathcal{L}$; the second term $V_{div}(\boldsymbol{f}, \mathcal{D})$ corresponds to the *diversity loss* of $\boldsymbol{f}$ on a specified data set $\mathcal{D}$ (e.g. $\mathcal{D} = \mathcal{U}$). Furthermore, γ is the cost parameter balancing the importance of the two terms.

In this paper, UDEED calculates the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ in Eq.(1) as:

$$V_{emp}(\boldsymbol{f}, \mathcal{L}) = \frac{1}{m} \cdot \sum_{k=1}^{m} l(f_k, \mathcal{L}) \qquad (2)$$

Here, $l(f_k, \mathcal{L})$ measures the empirical loss of the k-th base classifier f_k on the labeled data set $\mathcal{L}$.

As shown in Eq.(1), the second term $V_{div}(\boldsymbol{f}, \mathcal{D})$ is used to characterize the diversity among the based learners. However, it is well-known that diversity measurement is not a straightforward task since there is no generally accepted formal definition [12]. In this paper, UDEED chooses to calculate $V_{div}(\boldsymbol{f}, \mathcal{D})$ in a novel way as follows:

$$V_{div}(\boldsymbol{f}, \mathcal{D}) = \frac{2}{m(m-1)} \cdot \sum_{p=1}^{m-1} \sum_{q=p+1}^{m} d(f_p, f_q, \mathcal{D})$$

$$\text{where} \quad d(f_p, f_q, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} f_p(\boldsymbol{x}) f_q(\boldsymbol{x}) \tag{3}$$

Here, $|\mathcal{D}|$ returns the cardinality of data set $\mathcal{D}$. Intuitively, $d(f_p, f_q, \mathcal{D})$ represents the *prediction difference* between any pair of base classifiers on a specified data set $\mathcal{D}$.[1] In addition, the prediction difference is calculated based on the concrete output $f(\boldsymbol{x})$ instead of the signed output $\text{sign}[f(\boldsymbol{x})]$. In this way, the *prediction confidence* of each classifier other than the simple *binary prediction* is fully utilized.

Then, UDEED aims to find the target model $\boldsymbol{f}^*$ which minimizes the loss function in Eq.(1):

$$\boldsymbol{f}^* = \arg\min_{\boldsymbol{f}} V(\boldsymbol{f}, \mathcal{L}, \mathcal{D}) \tag{4}$$

B. Logistic Regression Implementation

In this paper, we employ *logistic regression* to implement the base classifiers. Specifically, each base classifier f_k ($1 \leq k \leq m$) is modeled as:

$$f_k(\boldsymbol{x}) = 2 \cdot g_k(\boldsymbol{x}) - 1 = 2 \cdot \frac{1}{1 + e^{-(\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x} + b_k)}} - 1 \tag{5}$$

Here, $g_k : \mathcal{X} \rightarrow [0, 1]$ is the standard logistic regression function with weight vector $\boldsymbol{w}_k \in \mathcal{R}^d$ and bias value $b_k \in \mathcal{R}$. Without loss of generality, in the rest of this paper, b_k is absorbed into $\boldsymbol{w}_k$ by appending the input space $\mathcal{X}$ with an extra dimension fixed at value 1.

Correspondingly, the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ in Eq.(1) is set to be the negative *binomial likelihood* function on the labeled data set $\mathcal{L}$, which is commonly used to measure the empirical loss of logistic regression:

$$\begin{aligned} V_{emp}(\boldsymbol{f}, \mathcal{L}) &= \frac{1}{m} \cdot \sum_{k=1}^{m} l(f_k, \mathcal{L}) \\ &= \frac{1}{mL} \cdot \sum_{k=1}^{m} \sum_{i=1}^{L} -\text{BLH}(f_k(\boldsymbol{x}_i), y_i) \end{aligned}$$

Here, the term $\text{BLH}(f_k(\boldsymbol{x}_i), y_i)$ calculates the binomial likelihood of $\boldsymbol{x}_i$ having label y_i, when f_k serves as the classification model. Note that the probabilities of $P(y = +1|\boldsymbol{x})$ and $P(y = -1|\boldsymbol{x})$ are modeled as $\frac{1+f_k(\boldsymbol{x})}{2}$ and $\frac{1-f_k(\boldsymbol{x})}{2}$ respectively, $\text{BLH}(f_k(\boldsymbol{x}_i), y_i)$ then takes the following form based on Eq.(5):

$$\begin{aligned} &\text{BLH}(f_k(\boldsymbol{x}_i), y_i) \\ &= \ln \left(\left(\frac{1 + f_k(\boldsymbol{x}_i)}{2} \right)^{\frac{1+y_i}{2}} \left(\frac{1 - f_k(\boldsymbol{x}_i)}{2} \right)^{\frac{1-y_i}{2}} \right) \\ &= -\frac{1+y_i}{2} \ln \left(1 + e^{-\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x}_i} \right) - \frac{1-y_i}{2} \ln \left(1 + e^{\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x}_i} \right) \end{aligned} \tag{6}$$

Note that the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ can also be evaluated in other ways, such as l_2 loss: $\frac{1}{mL} \sum_{k=1}^{m} \sum_{i=1}^{L} (f_k(\boldsymbol{x}_i) - y_i)^2$, hinge loss: $\frac{1}{mL} \sum_{k=1}^{m} \sum_{i=1}^{L} 1 - y_i f_k(\boldsymbol{x}_i)$, etc.

The target model $\boldsymbol{f}^*$ is found by employing *gradient descent*-based techniques. Accordingly, the gradients of $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ with respect to the model parameters $\boldsymbol{\Theta} = \{\boldsymbol{w}_k | 1 \leq k \leq m\}$ are determined as follows:[2]

$$\frac{\partial V}{\partial \boldsymbol{\Theta}} = \left[\frac{\partial V}{\partial \boldsymbol{w}_1}, \cdots, \frac{\partial V}{\partial \boldsymbol{w}_k}, \cdots, \frac{\partial V}{\partial \boldsymbol{w}_m} \right], \quad \text{where}$$

$$\begin{aligned} \frac{\partial V}{\partial \boldsymbol{w}_k} = &-\frac{1}{mL} \cdot \sum_{i=1}^{L} \frac{\partial \, \text{BLH}(f_k(\boldsymbol{x}_i), y_i)}{\partial \boldsymbol{w}_k} \\ &+ \frac{2\gamma}{m(m-1)} \cdot \sum_{k'=1,\ k' \neq k}^{m} \frac{\partial \, d(f_k, f_{k'}, \mathcal{D})}{\partial \boldsymbol{w}_k}, \quad \text{and} \end{aligned}$$

$$\begin{aligned} &\frac{\partial \, \text{BLH}(f_k(\boldsymbol{x}_i), y_i)}{\partial \boldsymbol{w}_k} = \\ &\left(\frac{(1 + y_i)(1 - f_k(\boldsymbol{x}_i))}{4} \ln \left(1 + e^{-\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x}_i} \right) \right. \\ &\left. - \frac{(1 - y_i)(1 + f_k(\boldsymbol{x}_i))}{4} \ln \left(1 + e^{\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x}_i} \right) \right) \cdot \boldsymbol{x}_i, \quad \text{and} \end{aligned}$$

$$\frac{\partial \, d(f_k, f_{k'}, \mathcal{D})}{\partial \boldsymbol{w}_k} = \frac{1}{2|\mathcal{D}|} \cdot \sum_{\boldsymbol{x} \in \mathcal{D}} f_{k'}(\boldsymbol{x}) \cdot (1 - f_k(\boldsymbol{x})^2) \cdot \boldsymbol{x} \tag{7}$$

To initialize the ensemble, each classifier f_k is learned from a *bootstrapped sample* of $\mathcal{L}$, namely $\mathcal{L}_k = \{(\boldsymbol{x}_i^k, y_i^k) | 1 \leq i \leq L\}$, by conventional maximum likelihood procedure. Specifically, the corresponding model parameter $\boldsymbol{w}_k$ is obtained by minimizing the objective function $\frac{1}{2}||\boldsymbol{w}_k||^2 + \lambda \cdot \sum_{i=1}^{L} -\text{BLH}(f_k(\boldsymbol{x}_i^k), y_i^k)$. Here, λ balances the model complexity and the binomial likelihood of f_k on $\mathcal{L}_k$. In

[1] As reviewed in [12], most existing diversity measures are calculated based on the *oracle* (correct/incorrect) outputs of base learners, i.e. the *ground-truth* labels of the data set are assumed to be known. However, considering that examples contained in the specified data set $\mathcal{D}$ may be *unlabeled*, it is then infeasible to calculate $d(f_p, f_q, \mathcal{D})$ by directly utilizing existing diversity measures.

[2] Note that under logistic regression implementation, the loss function $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ is generally *non-convex*, and the target model $\boldsymbol{f}^*$ returned by the gradient descent process would correspond to a *local* optimal solution.

Table I
CHARACTERISTICS OF THE DATA SETS (d: DIMENSIONALITY, *pos.*: #POSITIVE EXAMPLES, *neg.*: #NEGATIVE EXAMPLES).

data set	d	*pos./neg.*	data set	d	*pos./neg.*	data set	d	*pos./neg.*	data set	d	*pos./neg.*	data set	d	*pos./neg.*
diabetes	8	268/500	vote	16	168/267	ionosphere	34	255/126	credit_g	61	300/700	adult	123	7841/24720
heart	9	120/150	vehicle	16	218/217	kr_vs_kp	40	1527/1669	BCI	117	200/200	web	300	1479/48270
wdbc	14	357/212	hepatitis	19	123/32	isolet	51	300/300	Digit1	241	734/766	ijcnn1	22	13565/128126
austra	15	307/383	labor	26	37/20	sonar	60	111/97	COIL2	241	750/750	cod-rna	8	110384/220768
house	16	108/124	ethn	30	1310/1320	colic	60	136/232	g241n	241	748/752	forest	54	283301/297711

this paper, λ is set to the default value of 1. Note that the ensemble can also be initialized in other ways, such as instantiating each $\boldsymbol{w}_k$ with random values, etc.

As shown in Eq.(1), the second term $V_{div}(\boldsymbol{f}, \mathcal{D})$ regarding ensemble diversity is defined on a specified data set $\mathcal{D}$. Given the labeled training set $\mathcal{L}$ and the unlabeled training set $\mathcal{U}$, we consider three possibilities of instantiating $\mathcal{D}$:

- $\mathcal{D} = \emptyset$: No data is employed to measure the diversity among base learners ($V_{div}(\boldsymbol{f}, \mathcal{D})$=0). The resulting implementation is called LC;
- $\mathcal{D} = \tilde{\mathcal{L}}$: Labeled training examples are employed to measure the diversity among base learners, and the ensemble is optimized by exploiting only $\mathcal{L}$. The resulting implementation is called LCD;
- $\mathcal{D} = \mathcal{U}$: Unlabeled training examples are employed to measure the diversity among base learners, and the ensemble is optimized by exploiting both $\mathcal{L}$ and $\mathcal{U}$. The resulting implementation is called LCUD;

For LC and LCD, after the ensemble is initialized, a series of *gradient descent* steps are performed to optimize the model by minimizing the loss function $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ as defined in Eq.(1). For LCUD however, instead of directly minimizing $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ in the straightforward way of setting $\mathcal{D} = \mathcal{U}$, the loss function is firstly minimized by a series of gradient descent steps with $\mathcal{D} = \tilde{\mathcal{L}}$. After that, by using the learned model as the *starting point*, a series of gradient descent steps are further conducted to finely search the model space with $\mathcal{D} = \mathcal{U}$. The purpose of this two-stage process is to distinguish the *priorities* of the contribution from labeled data and unlabeled data.[3]

For any *gradient descent*-based optimization process, it is terminated if either the loss function $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ or the diversity term $V_{div}(\boldsymbol{f}, \mathcal{D})$ does not decrease anymore. For each implementation of UDEED, the label of an unseen example $\boldsymbol{z}$ is predicted by the learned ensemble $\boldsymbol{f}^* = (f_1^*, f_2^*, \cdots, f_m^*)$ via *weighted voting*:[4] $\boldsymbol{f}^*(\boldsymbol{z}) = \text{sign}\left[\sum_{k=1}^{m} f_k^*(\boldsymbol{z})\right]$.

Intuitively, if the ensemble does benefit from the diversity augmented by the unlabeled training examples, LCUD should achieve superior performance than LC and LCD.

IV. EXPERIMENTS

In this section, comparative studies between UDEED (i.e. LCUD) and other semi-supervised ensemble methods are firstly reported. More importantly, experimental analysis on the three different implementations of UDEED are further conducted to show whether unlabeled data do benefit ensemble learning by helping augment the diversity among base learners.

Twenty-five publicly-available binary data sets are used for experiments, whose characteristics are summarized in Table I. Fifteen of them are from UCI Machine Learning Repository [2], five from UCI KDD Archive [10], four from [5] and one from [15]. Twenty *regular-scale* data sets (left four columns) as well as five *large-scale* data sets (right column) are included. The data set size varies from 57 to 581,012, the dimensionality varies from 8 to 300, and the ratio between positive examples to negative examples varies from 0.031 to 3.844.

For each data set, 50% of them are randomly selected to form the test set $\mathcal{T}$, and the rest is used to form the training set of $\mathcal{L} \bigcup \mathcal{U}$. The percentage of labeled data in training set (i.e. $|\mathcal{L}|/(|\mathcal{L}| + |\mathcal{U}|)$) is set to be 0.25. For each data set, 50 random $\mathcal{L}/\mathcal{U}/\mathcal{T}$ splits are performed. Hereafter, the reported performance of each method corresponds to the average result out of 50 runs on different splits.

Various ensemble sizes (i.e. m) are considered in the experiments: a) $m = 20$ representing the case of *small-scale* ensemble; b) $m = 50$ representing the case of *medium-scale* ensemble; and c) $m = 100$ representing the case of *large-scale* ensemble.[5] In addition, as shown in Eq.(1), the cost

[3]Similar strategies have been adopted by some successful semi-supervised ensemble methods [16], [18], where objective terms involving labeled data are given much higher weight than those involving unlabeled data.

[4]Compared to *unweighted voting* where the label of $\boldsymbol{z}$ is predicted by $\boldsymbol{f}^*(\boldsymbol{z}) = \text{sign}\left[\sum_{k=1}^{m} \text{sign}[f_k^*(\boldsymbol{z})]\right]$, the *prediction confidence* of each base learner could be fully utilized by weighted voting.

[5]Preliminary experiments show that, as the ensemble size increases from 10 to 100 within an interval of 100, the performance of UDEED does not significantly change within successive ensemble sizes and tends to converge as the ensemble size approaches 100.

Table II
PREDICTIVE ACCURACY (MEAN±STD.) UNDER *small-scale* ENSEMBLE SIZE ($m = 20$). •/○ INDICATES WHETHER UDEED IS STATISTICALLY SUPERIOR/INFERIOR TO THE COMPARED ALGORITHM (PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL).

	Algorithm				
Data Set	UDEED	BAGGING	ADABOOST	ASSEMBLE	SEMIBOOST
diabetes	0.726±0.021	0.690±0.018•	0.728±0.029	0.700±0.031•	0.695±0.019•
heart	0.793±0.040	0.779±0.043•	0.766±0.045•	0.744±0.072•	0.789±0.035
wdbc	0.927±0.014	0.807±0.024•	0.934±0.025	0.898±0.070•	0.793±0.028•
austra	0.834±0.023	0.810±0.024•	0.809±0.028•	0.801±0.038•	0.815±0.029•
house	0.921±0.028	0.922±0.027	0.849±0.156•	0.921±0.036	0.924±0.029
vote	0.932±0.017	0.930±0.018•	0.906±0.106	0.928±0.019	0.932±0.017
vehicle	0.916±0.019	0.914±0.021	0.916±0.064	0.921±0.029	0.886±0.026•
hepatitis	0.800±0.042	0.792±0.026	0.763±0.077•	0.788±0.041	0.796±0.026
labor	0.809±0.072	0.801±0.074	0.646±0.142•	0.747±0.075•	0.810±0.071
ethn	0.944±0.007	0.942±0.008•	0.934±0.013•	0.939±0.010•	0.929±0.009•
ionosphere	0.795±0.043	0.721±0.023•	0.807±0.037	0.772±0.038•	0.746±0.027•
kr_vs_kp	0.940±0.008	0.938±0.008•	0.941±0.009	0.942±0.010	0.936±0.008•
isolet	0.989±0.007	0.988±0.006	0.714±0.244•	0.985±0.010•	0.989±0.005
sonar	0.690±0.069	0.690±0.070	0.701±0.063	0.672±0.068	0.692±0.067
colic	0.777±0.035	0.785±0.035○	0.747±0.039•	0.748±0.037•	0.765±0.041•
credit_g	0.690±0.024	0.710±0.019○	0.678±0.023•	0.686±0.025	0.702±0.019○
BCI	0.582±0.039	0.576±0.039•	0.606±0.040○	0.575±0.037	0.569±0.049•
Digit1	0.939±0.010	0.940±0.009	0.928±0.012•	0.927±0.012•	0.941±0.009○
COIL2	0.807±0.029	0.809±0.028	0.862±0.017○	0.819±0.023○	0.823±0.021○
g241n	0.793±0.020	0.794±0.018	0.760±0.021•	0.751±0.020•	0.791±0.022
adult	0.835±0.003	0.844±0.002○	0.840±0.003○	0.843±0.002○	N/A
web	0.981±0.001	0.980±0.001•	0.980±0.001•	0.981±0.001○	N/A
ijcnn1	0.914±0.001	0.906±0.001•	0.910±0.004•	0.906±0.001•	N/A
cod-rna	0.920±0.001	0.850±0.001•	0.945±0.003○	0.851±0.002•	N/A
forest	0.706±0.002	0.703±0.002•	0.736±0.006○	0.696±0.002•	N/A
win/tie/loss	／	**13/9/3**	**13/7/5**	**14/8/3**	**9/8/3**

parameter γ is set to the default value of 1. Note that better performance can be expected if certain strategies such as cross-validation are employed to optimize the value of γ.

A. Comparative Studies

In this subsection, UDEED (LCUD) is compared with two popular ensemble methods BAGGING [4] and ADABOOST [9], and two successful semi-supervised ensemble methods ASSEMBLE [1] and SEMIBOOST [16]. For fair comparison, logistic regression is employed as the base learner of each compared method. For UDEED, the maximum number of gradient descent steps is set to 25 and the learning rate is set to 0.25. For the other compared methods, default parameters suggested in respective literatures are adopted.

Tables II to IV report the detailed experimental results under *small-scale* (m=20), *medium-scale* (m=50) and *large-scale* (m=100) ensemble sizes respectively. SEMIBOOST fails to work on the *large-scale* data sets, due to its demanding storage complexity ($\mathcal{O}((|\mathcal{L}| + |\mathcal{U}|)^2)$) to maintain the similarity matrix for the training examples.

On each data set, the mean predictive accuracy as well as the standard deviation of each algorithm (out of 50 runs) are recorded. Furthermore, to statistically measure the significance of performance difference, pairwise t-tests at 95% significance level are conducted between the algorithms. Specifically, whenever UDEED achieves significantly better/worse performance than the compared algorithm on any data set, a win/loss is counted and a maker •/○ is shown. Otherwise, a tie is counted and no marker is given. The resulting win/tie/loss counts for UDEED against the compared algorithms are highlighted in the last line of each table.

In summary, when the ensemble size is *small* (Table II), UDEED is statistically superior to BAGGING, ADABOOST, ASSEMBLE and SEMIBOOST in 52%, 52%, 56% and 45% cases, and is inferior to them in much less 12%, 20%, 12% and 15% cases; When the ensemble size is *medium* (Table III), UDEED is statistically superior to BAGGING,

Table III
PREDICTIVE ACCURACY (MEAN±STD.) UNDER *medium-scale* ENSEMBLE SIZE ($m = 50$). •/○ INDICATES WHETHER UDEED IS STATISTICALLY SUPERIOR/INFERIOR TO THE COMPARED ALGORITHM (PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL).

	Algorithm				
Data Set	UDEED	BAGGING	ADABOOST	ASSEMBLE	SEMIBOOST
diabetes	0.710±0.020	0.691±0.019•	0.731±0.026○	0.699±0.032•	0.696±0.019•
heart	0.794±0.033	0.782±0.032•	0.766±0.037•	0.736±0.078•	0.794±0.033
wdbc	0.885±0.017	0.806±0.022•	0.925±0.065○	0.916±0.046○	0.816±0.033•
austra	0.828±0.024	0.812±0.028•	0.808±0.025•	0.815±0.036•	0.816±0.029•
house	0.921±0.030	0.920±0.030	0.793±0.195•	0.925±0.034	0.924±0.029○
vote	0.931±0.017	0.929±0.018•	0.868±0.151•	0.927±0.019	0.932±0.017
vehicle	0.914±0.022	0.914±0.021	0.914±0.088	0.919±0.025	0.893±0.026•
hepatitis	0.796±0.031	0.792±0.022	0.737±0.106•	0.785±0.045	0.797±0.027
labor	0.813±0.083	0.799±0.079•	0.681±0.142•	0.749±0.095•	0.804±0.083
ethn	0.944±0.006	0.942±0.007•	0.937±0.013•	0.939±0.011•	0.931±0.009•
ionosphere	0.797±0.042	0.722±0.022•	0.814±0.035○	0.783±0.027•	0.748±0.028•
kr_vs_kp	0.939±0.008	0.938±0.008•	0.943±0.011○	0.943±0.009○	0.935±0.008•
isolet	0.989±0.006	0.988±0.007•	0.672±0.232•	0.986±0.008•	0.990±0.005
sonar	0.687±0.069	0.690±0.072	0.714±0.059○	0.679±0.070	0.696±0.068
colic	0.783±0.033	0.783±0.036	0.744±0.043•	0.748±0.046•	0.763±0.040•
credit_g	0.703±0.024	0.711±0.020○	0.674±0.026•	0.689±0.025•	0.703±0.019
BCI	0.582±0.041	0.577±0.041	0.620±0.043○	0.583±0.051	0.572±0.045•
Digit1	0.941±0.010	0.940±0.010	0.929±0.012•	0.925±0.012•	0.941±0.009
COIL2	0.808±0.027	0.812±0.024	0.867±0.016○	0.821±0.022○	0.820±0.022○
g241n	0.796±0.019	0.794±0.018	0.762±0.023•	0.750±0.020•	0.791±0.022•
adult	0.842±0.002	0.844±0.002○	0.841±0.002•	0.842±0.002○	N/A
web	0.981±0.001	0.980±0.001•	0.980±0.001	0.981±0.001○	N/A
ijcnn1	0.907±0.001	0.906±0.001•	0.906±0.001•	0.910±0.004○	N/A
cod-rna	0.891±0.001	0.851±0.001•	0.945±0.003○	0.851±0.003•	N/A
forest	0.705±0.002	0.703±0.002•	0.737±0.006○	0.698±0.003•	N/A
win/tie/loss	／	**14/9/2**	**14/2/9**	**13/6/6**	**10/8/2**

ADABOOST, ASSEMBLE and SEMIBOOST in 56%, 56%, 52% and 50% cases, and is inferior to them in much less 8%, 36%, 24% and 10% cases; When the ensemble size is *large* (Table IV), UDEED is statistically superior to BAGGING, ADABOOST, ASSEMBLE and SEMIBOOST in 48%, 52%, 52% and 40% cases, and is inferior to them in much less 8%, 40%, 20% and 15% cases. These results indicate that UDEED is highly competitive to the other compared methods. Roughly speaking, as for the time complexity, UDEED is slightly higher than BAGGING and ADABOOST while fairly comparable to ASSEMBLE and SEMIBOOST.

B. The Helpfulness of Unlabeled Data

As motivated in Section I, UDEED aims to exploit unlabeled data to help ensemble learning in the particular way of augmenting diversity among base learners. Therefore, in addition to the above comparative experiments with other (semi-supervised) ensemble methods, it is more important to show whether UDEED (LCUD) does achieve better performance than its counterparts (LC and LCD) which do not consider using unlabeled data for diversity augmentation.

Table V reports the performance improvement (i.e. increase of predictive accuracy) of LCUD against LC and LCD under various ensemble sizes. On each data set, the mean improved predictive accuracy as well as the standard deviation (out of 50 runs) are recorded. In addition, to statistically measure the significance of performance difference, pairwise t-tests at 95% significance level are conducted. Specifically, whenever LCUD achieves significantly superior/inferior performance than LC or LCD on any data set, a win/loss is counted and a maker •/○ is shown in the Table. Otherwise, a tie is counted and no marker is given. The resulting win/tie/loss counts for LCUD against LC and LCD are highlighted in the last line of Table V.

In summary, when the ensemble size is *small*, LCUD is statistically superior to LC and LCD in 64% and 56% cases, and is inferior to them in both only 12% cases; When the ensemble size is *medium*, LCUD is statistically superior to

Table IV
PREDICTIVE ACCURACY (MEAN±STD.) UNDER *large-scale* ENSEMBLE SIZE ($m = 100$). •/○ INDICATES WHETHER UDEED IS STATISTICALLY SUPERIOR/INFERIOR TO THE COMPARED ALGORITHM (PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL).

	Algorithm				
Data Set	UDEED	BAGGING	ADABOOST	ASSEMBLE	SEMIBOOST
diabetes	0.700±0.020	0.692±0.018•	0.726±0.032○	0.694±0.031	0.696±0.018•
heart	0.790±0.035	0.781±0.035•	0.757±0.041•	0.751±0.066•	0.792±0.036
wdbc	0.852±0.021	0.805±0.019•	0.930±0.064○	0.916±0.037○	0.825±0.030•
austra	0.824±0.025	0.812±0.024•	0.806±0.027•	0.808±0.038•	0.817±0.028•
house	0.921±0.028	0.921±0.029	0.831±0.180•	0.919±0.029	0.924±0.029○
vote	0.930±0.017	0.930±0.018	0.902±0.104	0.926±0.020	0.932±0.017○
vehicle	0.913±0.022	0.915±0.022	0.930±0.026○	0.911±0.031	0.897±0.027•
hepatitis	0.797±0.027	0.790±0.023•	0.743±0.101•	0.782±0.040•	0.797±0.026
labor	0.811±0.080	0.808±0.080	0.683±0.146•	0.756±0.098•	0.809±0.075
ethn	0.943±0.007	0.942±0.007	0.938±0.012•	0.939±0.011•	0.932±0.008•
ionosphere	0.780±0.032	0.721±0.023•	0.812±0.037○	0.779±0.042	0.747±0.027•
kr_vs_kp	0.939±0.008	0.938±0.007•	0.945±0.011○	0.944±0.008○	0.935±0.008•
isolet	0.989±0.006	0.989±0.006•	0.616±0.208•	0.984±0.012•	0.990±0.005
sonar	0.690±0.071	0.689±0.070	0.713±0.061○	0.679±0.063	0.696±0.069
colic	0.784±0.033	0.786±0.033	0.741±0.041•	0.745±0.051•	0.763±0.042•
credit_g	0.706±0.021	0.711±0.021○	0.679±0.024•	0.686±0.026•	0.703±0.019
BCI	0.580±0.041	0.578±0.042	0.620±0.043○	0.588±0.041	0.572±0.046
Digit1	0.940±0.009	0.940±0.010	0.927±0.013•	0.925±0.011•	0.941±0.009
COIL2	0.807±0.027	0.811±0.024	0.870±0.016○	0.819±0.027○	0.820±0.021○
g241n	0.795±0.018	0.796±0.018	0.760±0.023•	0.754±0.027•	0.792±0.022
adult	0.844±0.002	0.844±0.002○	0.840±0.002•	0.843±0.002•	N/A
web	0.981±0.001	0.980±0.001•	0.980±0.002	0.981±0.001○	N/A
ijcnn1	0.906±0.001	0.905±0.004•	0.906±0.001•	0.906±0.001○	N/A
cod-rna	0.873±0.001	0.851±0.001•	0.945±0.003○	0.851±0.003•	N/A
forest	0.705±0.002	0.703±0.002•	0.737±0.006○	0.698±0.003•	N/A
win/tie/loss	／	**12/11/2**	**13/2/10**	**13/7/5**	**8/9/3**

LC and LCD in both 52% cases, and is inferior to them in both only 8% cases; When the ensemble size is *large*, LCUD is statistically superior to LC and LCD in 52% and 56% cases, and is inferior to them in only 8% and 12% cases. These results indicate that, by exploiting unlabeled data in the specific way of helping augment ensemble diversity, UDEED (LCUD) is capable of achieving better performance than its counterparts (LC and LCD) which do not consider employing unlabeled in ensemble generation.[6]

C. Diversity Analysis

To clearly verify that UDEED (LCUD) does increase the diversity among base learners after generating ensemble by utilizing unlabeled data, additional experiments are analyzed in this subsection based on several existing diversity measures. Specifically, four diversity measures summarized in [12] are considered, whose values are calculated based on the *oracle* (correct/incorrect) outputs of base learners.

Suppose m denotes the number of base classifiers in the ensemble and N denotes the number of examples in the test set $\mathcal{T}$. In addition, let $\boldsymbol{O} = [o_{ij}]_{m\times N}$ be the oracle output matrix. Here, $o_{ij} = 1$ if the i-th base learner correctly classifies the j-th test example ($1 \le i \le m$, $1 \le j \le N$). Otherwise, $o_{ij} = 0$. The formal definitions of the four diversity measures are as follows:

• *Disagreement measure* (DIS):

$$\mathrm{DIS} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \mathrm{dis}_{ik}, \quad \text{where}$$

$$\mathrm{dis}_{ik} = \frac{\sum_{j=1}^{N} o_{ij} \cdot (1 - o_{kj}) + \sum_{j=1}^{N} (1 - o_{ij}) \cdot o_{kj}}{N}$$

[6]Note that although in a number of cases the accuracy difference between two algorithms looks rather marginal (e.g. less than 1%), the difference may still be statistically significant according to the pairwise t-test.

Table V
ACCURACY IMPROVEMENT (MEAN±STD.) FOR LCUD AGAINST LC AND LCD UNDER VARIOUS ENSEMBLE SIZES. •/○ INDICATES WHETHER LCUD IS STATISTICALLY SUPERIOR/INFERIOR TO THE COMPARED IMPLEMENTATION (PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL).

	Accuracy Improvement of LCUD against					
	LC			LCD		
Data Set	$m = 20$	$m = 50$	$m = 100$	$m = 20$	$m = 50$	$m = 100$
diabetes	0.034±0.024•	0.019±0.013•	0.008±0.011•	0.011±0.012•	0.009±0.009•	0.004±0.007•
heart	0.023±0.027•	0.009±0.016•	0.006±0.013•	0.009±0.016•	0.003±0.010•	0.004±0.009•
wdbc	0.127±0.024•	0.075±0.012•	0.047±0.013•	0.033±0.014•	0.031±0.013•	0.023±0.008•
austra	0.022±0.022•	0.015±0.013•	0.010±0.008•	0.004±0.012•	0.006±0.008•	0.005±0.005•
house	0.003±0.010•	-0.001±0.005	0.001±0.004•	0.002±0.007•	0.000±0.004	0.001±0.003•
vote	0.002±0.005•	0.001±0.003•	0.001±0.003•	0.001±0.004	0.001±0.002•	0.001±0.001•
vehicle	0.005±0.010•	0.002±0.005	0.001±0.004	0.003±0.007•	0.001±0.005	0.001±0.004
hepatitis	0.010±0.035	0.005±0.027	0.008±0.017•	0.003±0.027	0.001±0.019	0.005±0.012•
labor	0.003±0.071	0.004±0.043	0.004±0.018	-0.007±0.041	0.007±0.032	0.004±0.012•
ethn	0.002±0.003•	0.001±0.002•	0.001±0.002•	0.001±0.002•	0.001±0.001•	0.001±0.001•
ionosphere	0.073±0.049•	0.076±0.049•	0.057±0.035•	0.015±0.034•	0.022±0.032•	0.029±0.024•
kr_vs_kp	0.002±0.003•	0.001±0.002•	0.001±0.001•	0.001±0.001•	0.001±0.001•	0.001±0.001•
isolet	0.001±0.003•	0.001±0.002•	0.001±0.002	0.001±0.002	0.001±0.001•	0.001±0.001
sonar	0.001±0.036	0.003±0.022	0.001±0.015	0.002±0.016	-0.001±0.014	0.001±0.011
colic	-0.006±0.014○	-0.003±0.012	-0.001±0.008	-0.003±0.010○	-0.003±0.009	0.001±0.006
credit_g	-0.019±0.017○	-0.008±0.010○	-0.005±0.008○	-0.009±0.010○	-0.004±0.006○	-0.002±0.006○
BCI	0.006±0.015•	0.003±0.010	0.002±0.012	0.005±0.010•	0.002±0.010	0.002±0.011
Digit1	0.001±0.005	0.001±0.002	0.001±0.004	0.001±0.005	0.001±0.002	0.001±0.003
COIL2	-0.001±0.016	-0.004±0.016	-0.003±0.015	0.001±0.005	-0.001±0.006	-0.002±0.007○
g241n	0.001±0.005	0.001±0.004	-0.001±0.004	-0.001±0.004	0.001±0.004	-0.001±0.004
adult	-0.009±0.002○	-0.002±0.002○	-0.001±0.001○	-0.006±0.001○	-0.002±0.001○	-0.001±0.001○
web	0.001±0.001•	0.001±0.001•	0.000±0.000	0.001±0.001•	0.001±0.001•	0.000±0.000
ijcnn1	0.008±0.001•	0.001±0.001•	0.001±0.001•	0.006±0.001•	0.001±0.001•	0.001±0.001•
cod-rna	0.069±0.001•	0.041±0.001•	0.023±0.001•	0.022±0.001•	0.018±0.001•	0.011±0.001•
forest	0.003±0.001•	0.002±0.001•	0.001±0.001•	0.001±0.001•	0.001±0.001•	0.001±0.001•
win/tie/loss	**16/6/3**	**13/10/2**	**13/10/2**	**14/8/3**	**13/10/2**	**14/8/3**

• *Double-fault measure* (DF):

$$\mathrm{DF} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \mathrm{df}_{ik}, \quad \text{where}$$

$$\mathrm{df}_{ik} = \frac{\sum_{j=1}^{N} (1 - o_{ij}) \cdot (1 - o_{kj})}{N}$$

• *Entropy measure* (ENT):

$$\mathrm{ENT} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{m - \lceil m/2 \rceil} \min \left\{ \sum_{i=1}^{m} o_{ij}, m - \sum_{i=1}^{m} o_{ij} \right\}$$

• *Coincident failure diversity* (CFD):

$$\mathrm{CFD} = \begin{cases} 0, & p_0 = 1.0 \\ \frac{1}{1-p_0} \sum_{i=1}^{m} \frac{m-i}{m-1} p_i, & p_0 < 1.0 \end{cases}, \quad \text{where}$$

$$p_i = \frac{\sum_{j=1}^{N} \mathbf{1}_{\left[i = \sum_{k=1}^{m} (1 - o_{kj})\right]}}{N}, \quad (0 \le i \le m)$$

Here, DIS and DF are *pairwise* measures while ENT and CFD are *non-pairwise* measures. In addition, 1-DF is used instead of DF such that for all the measures, the *greater* the value the *higher* the diversity. All the four measures vary between 0 and 1.

Table VI compares UDEED's *initial* diversity after ensemble initialization with its *final* diversity after ensemble learning under various ensemble sizes. For each data set, pairwise t-tests at 95% significance level are conducted between the initial and the final ensemble diversities. Whenever the final

Table VI
THE WIN/TIE/LOSS RESULTS FOR FINAL ENSEMBLE AGAINST INITIAL ENSEMBLE IN TERMS OF THE FOUR DIVERSITY MEASURES UNDER VARIOUS ENSEMBLE SIZES.

	FINAL ensemble vs. INITIAL ensemble											
	$m = 20$				$m = 50$				$m = 100$			
Data Set	DIS	DF	ENT	CFD	DIS	DF	ENT	CFD	DIS	DF	ENT	CFD
diabetes	win	win	win	win	win	win	win	win	win	win	win	win
heart	loss	win	loss	tie	loss	win	loss	loss	loss	win	loss	loss
wdbc	tie	win	tie	tie	tie	tie	tie	tie	tie	win	tie	tie
austra	loss	win	loss	tie	loss	win	loss	tie	loss	win	loss	loss
house	win	win	win	win	win	win	win	win	win	win	win	win
vote	win	win	win	win	win	win	win	win	win	win	win	win
vehicle	tie	tie	tie	tie	loss	tie	tie	tie	win	tie	win	tie
hepatitis	win	tie	win	win	win	win	win	win	win	win	win	win
labor	tie	tie	tie	tie	win	win	win	tie	win	win	win	tie
ethn	win	win	win	win	loss	tie	tie	tie	win	win	win	tie
ionosphere	win	win	win	win	win	win	win	win	win	win	win	win
kr_vs_kp	win	win	win	win	win	win	win	win	win	win	win	win
isolet	win	tie	win	tie	win	loss	win	tie	win	loss	win	tie
sonar	loss	tie	loss	loss	loss	tie	loss	tie	loss	tie	loss	tie
colic	win	loss	win	win	win	tie	win	tie	win	tie	win	tie
credit_g	win	loss	win	win	win	loss	win	win	win	loss	win	win
BCI	win	win	win	win	win	win	win	win	win	win	win	win
Digit1	win	win	win	win	win	win	win	win	win	win	win	win
COIL2	win	win	win	win	tie	win	tie	win	tie	win	tie	win
g241n	tie	loss	tie	tie	tie	tie	tie	tie	tie	loss	tie	tie
adult	win	loss	win	win	win	loss	win	win	win	win	win	win
web	win	win	win	win	win	win	win	win	win	win	win	win
ijcnn1	loss	loss	loss	loss	loss	loss	loss	loss	loss	loss	loss	loss
cod-rna	tie	win	tie	win	tie	win	tie	tie	win	win	tie	tie
forest	tie	tie	tie	tie	tie	tie	tie	tie	tie	tie	tie	tie
win/tie/loss	**15/6/4**	**14/6/5**	**15/6/4**	**15/8/2**	**14/5/6**	**14/7/4**	**14/7/4**	**12/11/2**	**17/4/4**	**17/4/4**	**16/5/4**	**12/10/3**

ensemble achieves significantly higher/lower diversity than the initial one, a win/loss is recorded. Otherwise, a tie is recorded. The resulting win/tie/loss counts are highlighted in the last line of Table VI.

In summary, when the ensemble size is *small*, UDEED statistically increases the initial ensemble diversity in 60% (DIS), 56% (DF), 60% (ENT) and 60% (CFD) cases, but decreases the initial ensemble diversity in only 16% (DIS), 20% (DF), 16% (ENT) and 8% (CFD) cases.

When the ensemble size is *medium*, UDEED statistically increases the initial ensemble diversity in 56% (DIS), 56% (DF), 56% (ENT) and 48% (CFD) cases, but decreases the initial ensemble diversity in only 24% (DIS), 16% (DF), 16% (ENT) and 8% (CFD) cases;

Finally, when the ensemble size is *large*, UDEED statistically increases the initial ensemble diversity in 68% (DIS), 68% (DF), 64% (ENT) and 48% (CFD) cases, but decreases the initial ensemble diversity in only 16% (DIS), 16% (DF), 16% (ENT) and 12% (CFD) cases.

These results clearly verify that UDEED can effectively exploit unlabeled data to help augment ensemble diversity.

V. CONCLUSION

Previous ensemble methods try to obtain a high accuracy of base learners and high diversity among base learners by considering only labeled data. There were some studies on using unlabeled data, but focusing on using unlabeled data to improve accuracy. The major contribution of our work is to use unlabeled data to augment diversity, which suggests a new direction for ensemble design. Specifically, a novel semi-supervised ensemble method named UDEED is proposed, which works by maximizing accuracy on labeled

data while maximizing diversity on unlabeled data.

Experiments show that: a) UDEED achieves highly comparable performance against other successful semi-supervised ensemble methods; b) UDEED does benefit from unlabeled data by using them to augment the diversity among base learners. In the future, it is interesting to see whether UDEED works well with other base learners. It would be insightful to analyze why UDEED can achieve good performance theoretically. Furthermore, designing other ensemble methods by exploiting unlabeled data to augment ensemble diversity gracefully is a direction very worth studying.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their helpful comments in improving this paper. This work was supported by the National Science Foundation of China (60635030, 60805022), the National Fundamental Research Program of China (2010CB327903), the Ph.D. Programs Foundation of Ministry of Education of China (200802941009), the Jiangsu Science Foundation (BK2008018) and the Jiangsu 333 Program.

REFERENCES

[1] K. Bennett, A. Demiriz, and R. Maclin, "Exploiting unlabeled data in ensemble methods," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 289–296.

[2] C. Blake, E. Keogh, and C. J. Merz, "UCI repository of machine learning databases [http://www.ics.uci.edu/˜mlearn/mlrepository.html]," Department of Information and Computer Science, University of California, Irvine, CA, Tech. Rep., 1998.

[3] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, WI, 1998, pp. 92–100.

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

[5] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[6] K. Chen and S. Wang, "Regularized boost for semi-supervised learning," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 281–288.

[7] F. d'Alché Buc, Y. Grandvalet, and C. Ambroise, "Semi-supervised marginboost," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 553–560.

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

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Lecture Notes in Computer Science 904*, P. M. B. Vitányi, Ed. Berlin: Springer, 1995, pp. 23–37.

[10] S. Hettich and S. D. Bay, "The UCI KDD archive [http://kdd.ics.uci.edu]," Department of Information and Computer Science, University of California, Irvine, CA, Tech. Rep., 1998.

[11] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1995, pp. 231–238.

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

[13] M. Li and Z.-H. Zhou, "SETRED: Self-training with editing," in *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data mining*, Hanoi, Vietnam, 2005, pp. 611–621.

[14] ——, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 6, pp. 1088–1098, 2007.

[15] X. Lu and A. K. Jain, "Ethnicity identification from face images," in *Proceedings of SPIE International Symposium on Defense and Security*, Kissimmee, FL, 2004, pp. 114–123.

[16] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "Semiboost: Boosting for semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2000–2014, 2009.

[17] L. Mason, P. Bartlett, J. Baxter, and M. Frean, "Functional gradient techniques for combining hypotheses," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 221–246.

[18] H. Valizadegan, R. Jin, and A. K. Jain, "Semi-supervised boosting for multi-class classification," in *Proceedings of the 19th European Conference on Machine Learning*, Antwerp, Belgium, 2008, pp. 522–537.

[19] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[20] Z.-H. Zhou, "When semi-supervised learning meets ensemble learning," in *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, Reykjavik, Iceland, 2009, pp. 529–538.

[21] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.

[22] ——, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.