Noname manuscript No.
(will be inserted by the editor)

Exploiting Unlabeled Data to Enhance Ensemble Diversity

Min-Ling Zhang · Zhi-Hua Zhou

Abstract Ensemble learning learns from the training data by generating an ensemble of multiple base learners. It is well-known that to construct a good ensemble with strong generalization ability, the base learners are deemed to 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, i.e. Unlabeled Data to Enhance Ensemble Diversity, is proposed. In contrast to existing semi-supervised ensemble methods which utilize unlabeled data by estimating error-prone *pseudo-labels* on them to enlarge the labeled data to improve base learners' accuracies, UDEED works by maximizing accuracies of base learners on labeled data while maximizing *diversity* among them on unlabeled data. Extensive experiments on twenty regular-scale and five large-scale data sets are conducted under the setting of either *few* or *abundant* labeled data. Experimental results show that UDEED can effectively utilize unlabeled data for ensemble learning via diversity augmentation, and is highly competitive to well-established semi-supervised ensemble methods.

Keywords Machine learning · Ensemble learning · Unlabeled data · Diversity

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

Zhi-Hua Zhou[†] (Corresponding Author)
[†]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
E-mail: zhouzh@lamda.nju.edu.cn

This research was supported by the National Science Foundation of China (61073097, 60805022, 61175049), the National Fundamental Research Program of China (2010CB327903), Ph.D. Programs Foundation of Ministry of Education of China for Young Faculties (200802941009), and the Cultivation Program for Young Faculties of Southeast University.

1 Introduction

In ensemble learning, a number of base learners are trained and then combined for prediction to achieve strong generalization ability (Dietterich 2000; Zhou 2009a). During the past two decades, a number of popular ensemble learning methods have been proposed, such as BOOSTING (Freund and Schapire 1995), BAGGING (Breiman 1996), STACKING (Wolpert 1992), etc. Most of these methods work under the supervised setting where the labels of training examples are assumed to be known. In many real-world tasks, however, unlabeled training examples are readily available while obtaining their labels would be fairly expensive due to the involved time, human and economic costs. Semi-supervised learning is one of the major paradigms to exploit unlabeled data together with labeled data to improve learning performance automatically, without human interventions (Chapelle et al 2006; Zhou and Li 2010; Zhu 2006).

This paper deals with semi-supervised ensembles, i.e. *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 works have been devoted to the study of semi-supervised ensembles. As recently indicated by Zhou (2009b), 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 (2009b) indicated, 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 (Krogh and Vedelsby 1995). 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 which extends our previous research on ensemble learning with labeled and unlabeled data (Zhang and Zhou 2010). Specifically, UDEED aims to maximize accuracies of base learners on labeled data while maximize diversity among them on unlabeled data. Extensive experiments over twenty-five data sets are conducted with either few or abundant labeled training data. Experimental results show that: a) By using unlabeled data for diversity augmentation, UDEED achieves much better performance than its counterpart which does not consider the usefulness of unlabeled data; and b) UDEED also achieves highly comparable performance to other state-of-the-art semi-supervised ensemble methods.

The rest of this paper is organized as follows. Section 2 briefly reviews related work on semi-supervised ensembles. Section 3 presents UDEED. Section 4 reports our experimental results. Section 5 discusses several related issues. Finally, Section 6 concludes.

2 Related Work

In ensemble learning area, there have been many works on building strong ensembles by trying to maximize ensemble diversity, such as imposing negative correlation constraints among base learners (Liu and Yao 1999a,b), utilizing artificial examples to encourage diversity (Melville and Mooney 2003; Melville 2005), etc. However, most of them are supervised ensembles which learn from labeled examples without considering unlabeled data. In this section, we will focus on reviewing related works on semi-supervised ensembles. 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 (2005) 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 (Zhou and Li 2010), which can be viewed as a variant of the famous *co-training* method (Blum and Mitchell 1998). Later, Li and Zhou (2007) 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 (Bennett et al 2002; d'Alché Buc et al 2002; Chen and Wang 2008; Mallapragada et al 2009; Valizadegan et al 2008). d'Alché Buc et al (2002) proposed SSMBOOST to handle unlabeled data within the margin cost functional optimization framework for boosting (Mason et al 2000), where the margin of an ensemble H on unlabeled data $\boldsymbol{x}$ is defined as either $H(\boldsymbol{x})^2$ or $|H(\boldsymbol{x})|$ with $H(\boldsymbol{x}) \in [-1, 1]$ being the ensemble output. Furthermore, SSMBOOST enforces that the base learners in the ensemble should be semi-supervised in their nature. Later, Bennett et al (2002) 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 learner which is then added to H. Following the same margin cost functional optimization framework, Chen and Wang (2008) regularized ASSEMBLE with local smoothness constraints to help induce new base learners with more reliable self-labeling processes.

Other than the margin cost functional formalization, MCSSB (Valizadegan et al 2008) and SEMIBOOST (Mallapragada et al 2009) 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. Here, MCSSB implemented the objective terms based on Bregman divergence (Valizadegan et al 2008), while SEMIBOOST implemented them with traditional exponential loss (Mallapragada et al 2009). Recently, regularization techniques have been introduced to semi-supervised boosting by exploiting information-theoretic principles (Saffari et al 2008, 2009) or multiple semi-supervised assumptions (Chen and Wang 2011) .

A commonness of the above semi-supervised ensemble methods is that they construct the ensembles iteratively, and in particular, 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 (Bennett et al 2002; d'Alché Buc et al 2002; Li and Zhou 2007; Zhou and Li 2005), or with specific form of smoothness or manifold regularization (Chen and Wang 2008; Mallapragada et al 2009; Valizadegan et al 2008). After that, by regarding the estimated labels as their *ground-truth* labels, unlabeled instances are used in conjunction with labeled ones 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 (Li and Zhou 2005; Zhang and Zhou in press), the estimated pseudo-labels may still be prone to error, especially in initial training iterations where the ensemble is less accurate. In the next section we will present the UDEED approach. Rather than assigning pseudo-labels on unlabeled data to enlarge labeled training set, UDEED utilizes

unlabeled data in a different way, i.e., help augment the *diversity* among base learners.

3 The UDEED Approach

3.1 General Formulation

Let $\mathcal{X} = \mathcal{R}^d$ be the d-dimensional real-valued input space and $\mathcal{Y} = \{-1, +1\}$ be the binary output space. Let $\mathcal{L} = \{(\boldsymbol{x}_i, y_i) | \ 1 \leq i \leq L\}$ be the *labeled training set* which contains L labeled training examples with $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, and $\mathcal{U} = \{\boldsymbol{x}_i | L+1 \leq i \leq L+U\}$ be the *unlabeled training set* which contains U unlabeled training examples with $\boldsymbol{x}_i \in \mathcal{X}$.

We assume that the classifier ensemble is composed of m base classifiers $\{f_k | 1 \leq k \leq m\}$, where each of them maps from the instance space to $[-1, +1]$, i.e. $f_k : \mathcal{X} \to [-1, +1]$. Here, the 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 $P(y = +1|\boldsymbol{x})$.

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

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

Here, the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ corresponds to the *empirical loss* of the ensemble on the labeled data set $\mathcal{L}$; the second term $V_{div}(\boldsymbol{f}, \mathcal{D})$ corresponds to the *diversity loss* of the ensemble on a specified data set $\mathcal{D}$, such as $\mathcal{D} = \mathcal{U}$. Furthermore, γ is the cost parameter balancing the importance of the two loss terms.[1]

The first loss term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ in Eq.(1) is calculated as:

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

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 loss term $V_{div}(\boldsymbol{f}, \mathcal{D})$ is used to characterize the diversity among the based learners based on data set $\mathcal{D}$. However, it is well-known that diversity measurement is not a straightforward task since there is no generally accepted formal definition (Kuncheva and Whitaker 2003).

[1] Similar objective functions with the combination of an accuracy term and a diversity term have been investigated in ensemble learning, though under the supervised setting without considering unlabeled data (Opitz 1999; Opitz and Shavlik 1996).

Moreover, most of the existing diversity measures need to refer to the *ground-truth* labels of the data for diversity calculation, which are then not directly applicable here if $\mathcal{D}$ contains unlabeled data.

In this paper, UDEED chooses to calculate $V_{div}(\boldsymbol{f}, \mathcal{D})$ in a novel way as:

$$V_{div}(\boldsymbol{f}, \mathcal{D}) = \frac{2}{m(m-1)} \cdot \sum_{p=1}^{m-1} \sum_{q=p+1}^{m} d(f_p, f_q, \mathcal{D}), \text{ where}$$
$$d(f_p, f_q, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} f_p(\boldsymbol{x}) f_q(\boldsymbol{x}) \tag{3}$$

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 $\mathcal{D}$, without referring to the *ground-truth* labels of the data.[2] Furthermore, note that 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, and at the same time enables UDEED's objective (Eq.(1)) being a continuous function easier to be optimized.

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

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

3.2 Logistic Regression Implementation

In this paper, to instantiate the general formulation of UDEED given in the above subsection, *logistic regression* is employed to implement the base classifiers due to its effectiveness and simplicity. Concretely, 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, "T" represents the vector transpose. Furthermore, function $g_k : \mathcal{X} \to [0, 1]$ takes the canonical form of logistic regression $\frac{1}{1+e^{-\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x} - b_k}}$ with d-dimensional weight vector $\boldsymbol{w}_k$ and bias value b_k. 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.

[2] As reviewed in (Kuncheva and Whitaker 2003), 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.

Correspondingly, the first loss 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:

$$\begin{aligned} V_{emp}(\boldsymbol{f}, \mathcal{L}) &= \frac{1}{m} \cdot \sum_{k=1}^{m} l(f_k, \mathcal{L}) \\ &= \frac{1}{mL} \cdot \sum_{k=1}^{m} \sum_{i=1}^{L} -\mathrm{BLH}(f_k(\boldsymbol{x}_i), y_i) \end{aligned} \tag{6}$$

Here, the term $\mathrm{BLH}(f_k(\boldsymbol{x}_i), y_i)$ returns the binomial likelihood of $\boldsymbol{x}_i$ having label y_i, when f_k serves as the classification model. Considering that the posteriori probabilities of $P(y = +1|\boldsymbol{x})$ and $P(y = -1|\boldsymbol{x})$ can be calculated as $\frac{1+f_k(\boldsymbol{x})}{2}$ and $\frac{1-f_k(\boldsymbol{x})}{2}$ respectively with respect to the model f_k, $\mathrm{BLH}(f_k(\boldsymbol{x}_i), y_i)$ will then take the following form based on Eq.(5):

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

Note that the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ may also be evaluated in other ways, such as l_2 loss:

$$V_{emp}(\boldsymbol{f}, \mathcal{L}) = \frac{1}{mL} \sum_{k=1}^{m} \sum_{i=1}^{L} (f_k(\boldsymbol{x}_i) - y_i)^2$$

or hinge loss:

$$V_{emp}(\boldsymbol{f}, \mathcal{L}) = \frac{1}{mL} \sum_{k=1}^{m} \sum_{i=1}^{L} 1 - y_i f_k(\boldsymbol{x}_i)$$

and other possible forms. Based on Eqs.(5) and (6), the global loss function $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ is instantiated as:

$$\begin{aligned} & V(\boldsymbol{f}, \mathcal{L}, \mathcal{D}) \\ &= \frac{1}{mL} \sum_{k=1}^{m} \sum_{i=1}^{L} -\mathrm{BLH}(f_k(\boldsymbol{x}_i), y_i) + \gamma \cdot \frac{2}{m(m-1)} \cdot \sum_{p=1}^{m-1} \sum_{q=p+1}^{m} d(f_p, f_q, \mathcal{D}) \\ &= \frac{1}{mL} \sum_{k=1}^{m} \sum_{i=1}^{L} \frac{1 + y_i}{2} \ln \left(1 + e^{-\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x}_i} \right) + \frac{1 - y_i}{2} \ln \left(1 + e^{\boldsymbol{w}_k^{\mathrm{T}} \cdot \boldsymbol{x}_i} \right) \\ &\quad + \frac{2\gamma}{m(m-1)} \cdot \sum_{p=1}^{m-1} \sum_{q=p+1}^{m} \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \left(\frac{2}{1 + e^{-\boldsymbol{w}_p^{\mathrm{T}} \cdot \boldsymbol{x}}} - 1 \right) \cdot \left(\frac{2}{1 + e^{-\boldsymbol{w}_q^{\mathrm{T}} \cdot \boldsymbol{x}}} - 1 \right) \end{aligned}$$

The target model $\boldsymbol{f}^*$ is found by employing *gradient descent*-based techniques. 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.

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:

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

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

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

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

To initialize the ensemble, each classifier f_k is learned from a *bootstrapped sample* (Efron and Tibshirani 1993) of $\mathcal{L}$, i.e. by *sampling with replacement* from $\mathcal{L}$ to form the training set $\mathcal{L}_k = \{(\boldsymbol{x}_i^k, y_i^k) | 1 \leq i \leq L\}$. Conventional maximum likelihood procedure is used to initialize the model parameter $\boldsymbol{w}_k$ by minimizing the following objective function:

$$\frac{1}{2}||\boldsymbol{w}_k||^2 + \lambda \cdot \sum_{i=1}^{L} -\mathrm{BLH}(f_k(\boldsymbol{x}_i^k), y_i^k)$$

Here, λ balances the model complexity (first term) and the binomial likelihood of f_k on $\mathcal{L}_k$ (second term). In this paper, λ is set to the default value of 1.

Recall that in Eq.(1), the first term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ regarding empirical loss is defined on the labeled training set $\mathcal{L}$, while the second term $V_{div}(\boldsymbol{f}, \mathcal{D})$ regarding diversity loss 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 designating $\mathcal{D}$:

• $\mathcal{D} = \emptyset$: No data is employed to measure the diversity among base learners, i.e. $V_{div}(\boldsymbol{f}, \mathcal{D})$=0. The resulting implementation is called LC;

• $\mathcal{D} = \tilde{\mathcal{L}}$: Here, $\tilde{\mathcal{L}} = \{\boldsymbol{x}_i | 1 \leq i \leq L\}$ denotes the unlabeled data set derived from $\mathcal{L}$. In this case, 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.[3]

For either LC or LCD, after the ensemble is initialized, a series of *gradient descent* steps are performed to optimize the model by minimizing the global loss function $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ with respective configuration of $\mathcal{D}$. For LCUD however, instead of minimizing $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ in the straightforward way of directly setting $\mathcal{D} = \mathcal{U}$, the loss function is firstly minimized by a series of gradient descent steps by setting $\mathcal{D} = \tilde{\mathcal{L}}$. After that, by using the obtained intermediate model as the *starting point*, a series of gradient descent steps are further conducted to finely search the model space by setting $\mathcal{D} = \mathcal{U}$. The purpose of this two-stage process is to distinguish the *priorities* of the contribution from labeled data and unlabeled data. In other words, the labeled training examples are exploited with top priority to firstly induce the intermediate model for subsequent optimization. Note that similar strategies have been adopted by some successful semi-supervised ensemble methods, where objective terms involving labeled data are given much higher weight than those involving unlabeled data (Mallapragada et al 2009; Valizadegan et al 2008). More justifications on this specific optimization choice of LCUD are given in Subsection 5.1.

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

Intuitively, if the ensemble does benefit from the diversity augmented by the unlabeled training examples, LCUD should achieve superior performance than the other two implementations LC and LCD. In this paper, the third implementation (i.e. LCUD) is referred to as UDEED, and the other two implementations (i.e. LC and LCD) can be viewed as degenerated versions of UDEED.

4 Experiments

4.1 Experimental Setup

Twenty-five publicly-available binary data sets are used for experiments, whose characteristics are summarized in Table 1. Fifteen of them are from UCI Ma-

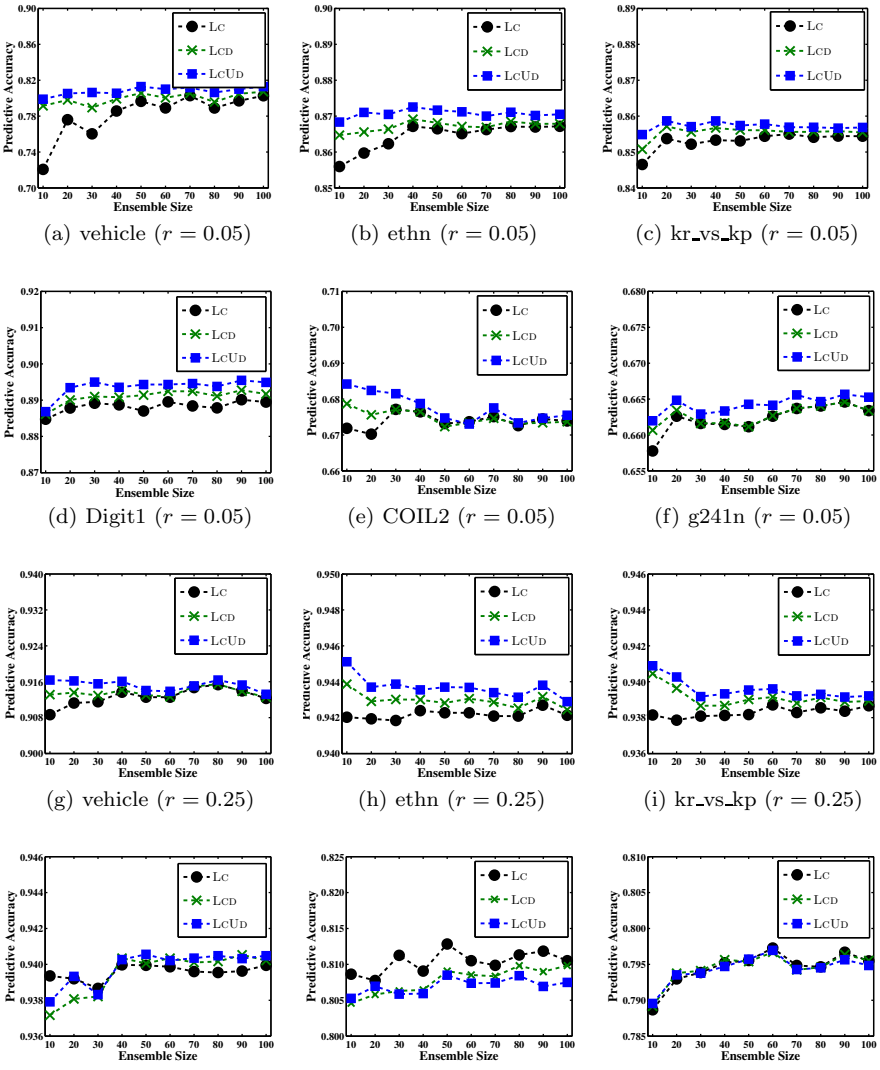
[3] The other possibility of designating $\mathcal{D}$ with $\mathcal{D} = \tilde{\mathcal{L}} \cup \mathcal{U}$ has also been considered, and the resulting implementation is called LCDUD. Preliminary experiments show that LCDUD performs rather similar as LCUD, and therefore only LCUD is studied in this paper for the simplicity of presentation.

Table 1 Characteristics of the data sets.

			# examples		
	data set	*dimensionality*	*positive*	*negative*	*total*
# regular-scale: 20	diabetes	8	268	500	768
	heart	9	120	150	270
	wdbc	14	357	212	569
	austra	15	307	383	690
	house	16	108	124	232
	vote	16	168	267	435
	vehicle	16	218	217	435
	hepatitis	19	123	32	155
	labor	26	37	20	57
	ethn	30	1310	1320	2630
	ionosphere	34	255	126	381
	kr_vs_kp	40	1527	1669	3196
	isolet	51	300	300	600
	sonar	60	111	97	208
	colic	60	136	232	368
	credit_g	61	300	700	1000
	BCI	117	200	200	400
	Digit1	241	734	766	1500
	COIL2	241	750	750	1500
	g241n	241	748	752	1500
# large-scale: 5	adult	123	7841	24720	32561
	web	300	1479	48270	49749
	ijcnn1	22	13565	128126	141691
	cod-rna	8	110384	220768	331152
	forest	54	283301	297711	581012

chine Learning Repository (Frank and Asuncion 2010), five from UCI KDD Archive (Hettich and Bay 1998), four from Chapelle et al (2006) and one from Lu and Jain (2004). Twenty *regular-scale* data sets (first part) as well as five *large-scale* data sets (second part) are included. Specifically, the data set size varies from 57 to 581,012, the dimensionality varies from 8 to 300, and the ratio of 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, i.e. $\mathcal{L} \bigcup \mathcal{U}$. Let $r = |\mathcal{L}|/(|\mathcal{L}| + |\mathcal{U}|)$ denote the percentage of labeled data in training set. 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 averaged result out of 50 runs on different splits. In this paper, r takes two different values: a) 0.05 representing the case that only *few* labeled data is available; b) 0.25 representing the case where



(a) vehicle ($r = 0.05$) (b) ethn ($r = 0.05$) (c) kr_vs_kp ($r = 0.05$)

(d) Digit1 ($r = 0.05$) (e) COIL2 ($r = 0.05$) (f) g241n ($r = 0.05$)

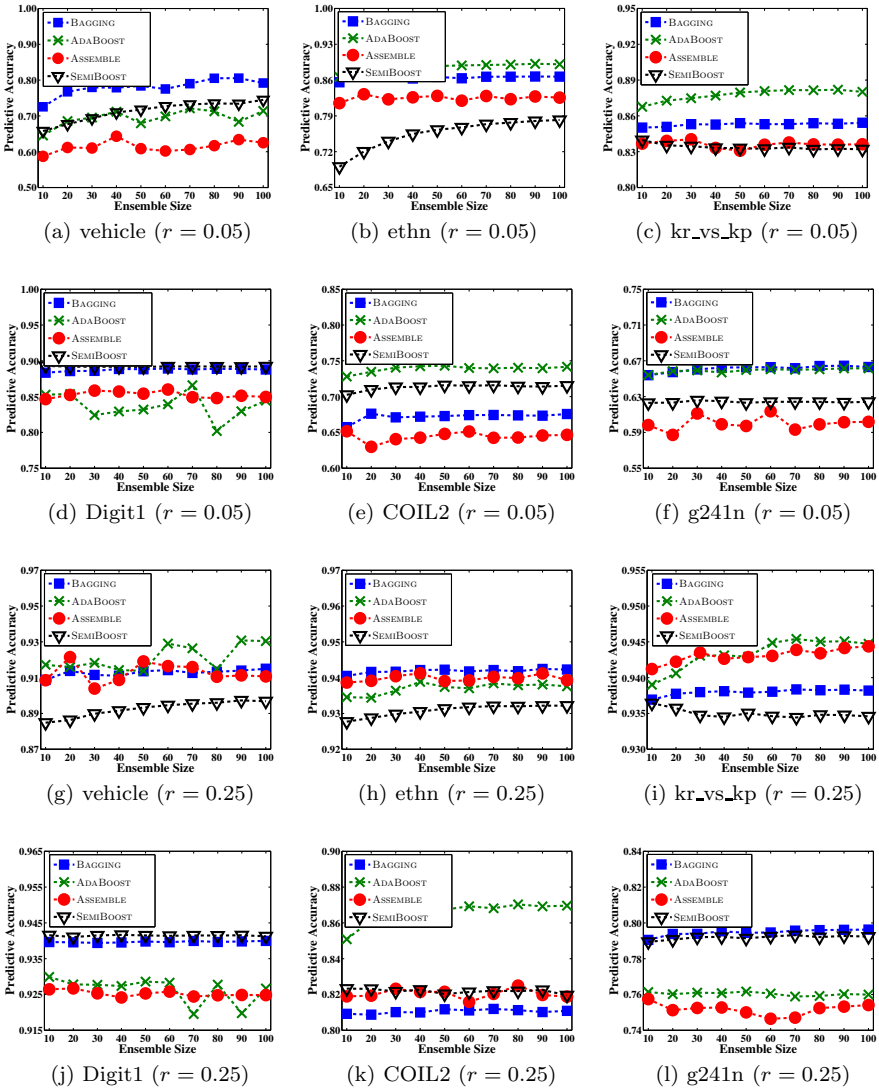
(g) vehicle ($r = 0.25$) (h) ethn ($r = 0.25$) (i) kr_vs_kp ($r = 0.25$)

(j) Digit1 ($r = 0.25$) (k) COIL2 ($r = 0.25$) (l) g241n ($r = 0.25$)

Fig. 1 Performance of Lc, Lcd and LcUd under varying ensemble sizes [(a)-(f): $r = 0.05$; (g)-(l): $r = 0.25$].

there are *abundant* labeled data.[4] As shown in Eq.(1), the cost parameter γ balances the empirical loss and diversity loss of the learning system. In this

[4] Generally, when there are only *very few* labeled data (say 2∼4 examples), it becomes infeasible to launch ensemble learning. While with the help of unlabeled data, it is possible to build ensembles under such situation (Zhou et al 2007), which is another advantage for using unlabeled data with ensembles (Zhou 2009b). However, this is beyond the scope of the paper since the goal of this paper is to show that *unlabeled data can help ensemble learning*



(a) vehicle ($r = 0.05$) (b) ethn ($r = 0.05$) (c) kr_vs_kp ($r = 0.05$)

(d) Digit1 ($r = 0.05$) (e) COIL2 ($r = 0.05$) (f) g241n ($r = 0.05$)

(g) vehicle ($r = 0.25$) (h) ethn ($r = 0.25$) (i) kr_vs_kp ($r = 0.25$)

(j) Digit1 ($r = 0.25$) (k) COIL2 ($r = 0.25$) (l) g241n ($r = 0.25$)

Fig. 2 Performance of BAGGING, ADABOOST, ASSEMBLE and SEMIBOOST under varying ensemble sizes [(a)-(f): $r = 0.05$; (g)-(l): $r = 0.25$].

paper, it is set to the default value of 1. More discussions on the choice of γ are given in Subsection 5.1.

In the rest of this paper, a series of experiments are conducted to validate the effectiveness of our proposed approach:

by enhancing diversity, instead of justifying the usefulness of unlabeled data for ensembles with sparse labeled data.

Table 2 Predictive accuracy of each comparing algorithm under *small-scale* ensemble size ($m = 20$). Bagging, AdaBoost, Assemble and SemiBoost are abbreviated as Bagg, AdaB, Assem and SemiB respectively.

	Algorithm									
	$r = 0.05$					$r = 0.25$				
Data Set	Udeed	Bagg	AdaB	Assem	SemiB	Udeed	Bagg	AdaB	Assem	SemiB
diabetes	0.670	0.652	0.656	0.648	**0.679**	0.726	0.690	**0.728**	0.700	0.695
heart	**0.670**	0.626	0.561	0.550	0.641	**0.793**	0.779	0.766	0.744	0.789
wdbc	**0.802**	0.646	0.700	0.646	0.661	0.927	0.807	**0.934**	0.898	0.793
austra	**0.714**	0.629	0.653	0.614	0.682	**0.834**	0.810	0.809	0.801	0.815
house	**0.889**	0.886	0.597	0.868	**0.889**	0.921	0.922	0.849	0.921	**0.924**
vote	0.898	**0.899**	0.658	0.894	0.896	**0.932**	0.930	0.906	0.928	**0.932**
vehicle	**0.805**	0.767	0.685	0.612	0.677	0.916	0.914	0.916	**0.921**	0.886
hepatitis	0.770	0.786	0.588	0.781	**0.795**	**0.800**	0.792	0.763	0.788	0.796
labor	**0.660**	0.650	0.500	0.570	0.613	0.809	0.801	0.646	0.747	**0.810**
ethn	0.871	0.867	**0.879**	0.832	0.720	**0.944**	0.942	0.934	0.939	0.929
ionosphere	**0.688**	0.664	0.663	0.667	0.677	0.795	0.721	**0.807**	0.772	0.746
kr_vs_kp	0.859	0.851	**0.873**	0.839	0.835	0.940	0.938	0.941	**0.942**	0.936
isolet	**0.964**	0.954	0.743	0.841	0.932	**0.989**	0.988	0.714	0.985	**0.989**
sonar	**0.575**	0.558	0.536	0.528	0.555	0.690	0.690	**0.701**	0.672	0.692
colic	0.714	**0.722**	0.654	0.663	0.691	0.777	**0.785**	0.747	0.748	0.765
credit_g	0.656	**0.695**	0.664	0.673	0.680	0.690	**0.710**	0.678	0.686	0.702
BCI	0.514	0.510	**0.515**	0.510	0.513	0.582	0.576	**0.606**	0.575	0.569
Digit1	**0.893**	0.886	0.854	0.852	0.892	0.939	0.940	0.928	0.927	**0.941**
COIL2	0.682	0.676	**0.735**	0.630	0.710	0.807	0.809	**0.862**	0.819	0.823
g241n	**0.665**	0.657	0.658	0.587	0.623	0.793	**0.794**	0.760	0.751	0.791
adult	0.831	**0.836**	0.818	0.831	N/A	0.835	**0.844**	0.840	0.843	N/A
web	**0.974**	0.973	0.971	**0.974**	N/A	**0.981**	0.980	0.980	**0.981**	N/A
ijcnn1	0.904	0.904	**0.918**	0.904	N/A	**0.914**	0.906	0.910	0.906	N/A
cod-rna	0.902	0.792	**0.933**	0.683	N/A	0.920	0.850	**0.945**	0.851	N/A
forest	0.707	0.703	**0.735**	0.688	N/A	0.706	0.703	**0.736**	0.696	N/A

- Firstly, comparative studies between Udeed (i.e. LcUd) and other well-established semi-supervised ensemble methods are reported;
- Secondly, experiments on the three different implementations Lc, Lcd and LcUd are further conducted to show whether unlabeled data do benefit ensemble learning;
- Thirdly, analysis based on several popular diversity measures is performed to verify whether the diversity among base learners can be enhanced by Udeed through utilizing unlabeled data in ensemble generation.

4.2 Comparative Studies

In this subsection, to evaluate the effectiveness of Udeed (LcUd) in ensemble learning, it is compared with two popular supervised ensemble methods Bagging (Breiman 1996) and AdaBoost (Freund and Schapire 1995), as

Table 3 Predictive accuracy of each comparing algorithm under *medium-scale* ensemble size ($m = 50$). BAGGING, ADABOOST, ASSEMBLE and SEMIBOOST are abbreviated as BAGG, ADAB, ASSEM and SEMIB respectively.

	Algorithm									
	$r = 0.05$					$r = 0.25$				
Data Set	UDEED	BAGG	ADAB	ASSEM	SEMIB	UDEED	BAGG	ADAB	ASSEM	SEMIB
diabetes	0.657	0.653	0.658	0.637	**0.684**	0.710	0.691	**0.731**	0.699	0.696
heart	**0.649**	0.612	0.555	0.554	0.642	**0.794**	0.782	0.766	0.736	**0.794**
wdbc	**0.710**	0.640	0.704	0.664	0.707	0.885	0.806	**0.925**	0.916	0.816
austra	0.685	0.622	0.654	0.593	**0.701**	**0.828**	0.812	0.808	0.815	0.816
house	0.889	0.890	0.680	0.833	**0.891**	0.921	0.920	0.793	**0.925**	0.924
vote	**0.899**	**0.899**	0.624	0.888	0.896	0.931	0.929	0.868	0.927	**0.932**
vehicle	**0.813**	0.784	0.679	0.609	0.717	0.914	0.914	0.914	**0.919**	0.893
hepatitis	0.778	0.786	0.640	0.779	**0.794**	0.796	0.792	0.737	0.785	**0.797**
labor	**0.663**	0.650	0.500	0.569	0.614	**0.813**	0.799	0.681	0.749	0.804
ethn	0.872	0.866	**0.888**	0.829	0.763	**0.944**	0.942	0.937	0.939	0.931
ionosphere	**0.699**	0.662	0.666	0.669	0.683	0.797	0.722	**0.814**	0.783	0.748
kr_vs_kp	0.857	0.854	**0.880**	0.831	0.833	0.939	0.938	**0.943**	**0.943**	0.935
isolet	**0.964**	0.959	0.724	0.870	0.931	0.989	0.988	0.672	0.986	**0.990**
sonar	**0.574**	0.569	0.537	0.546	0.552	0.687	0.690	**0.714**	0.679	0.696
colic	0.725	**0.732**	0.636	0.680	0.689	**0.783**	**0.783**	0.744	0.748	0.763
credit_g	0.677	**0.695**	0.652	0.682	0.680	0.703	**0.711**	0.674	0.689	0.703
BCI	**0.514**	0.511	0.510	0.502	0.512	0.582	0.577	**0.620**	0.583	0.572
Digit1	**0.894**	0.888	0.832	0.854	0.892	**0.941**	0.940	0.929	0.925	**0.941**
COIL2	0.675	0.673	**0.743**	0.648	0.716	0.808	0.812	**0.867**	0.821	0.820
g241n	**0.664**	0.663	0.660	0.597	0.623	**0.796**	0.794	0.762	0.750	0.791
adult	0.835	**0.837**	0.818	0.830	N/A	0.842	**0.844**	0.841	0.842	N/A
web	0.974	0.974	0.971	**0.975**	N/A	**0.981**	0.980	0.980	**0.981**	N/A
ijcnn1	0.904	0.904	**0.919**	0.904	N/A	0.907	0.906	0.906	**0.910**	N/A
cod-rna	0.855	0.793	**0.936**	0.683	N/A	0.891	0.851	**0.945**	0.851	N/A
forest	0.705	0.703	**0.735**	0.688	N/A	0.705	0.703	**0.737**	0.698	N/A

well as two successful semi-supervised ensemble methods ASSEMBLE (Bennett et al 2002) and SEMIBOOST (Mallapragada et al 2009). 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.

Figure 1 illustrates the performance curves of UDEED and its counterparts LC and LCD on six data sets with $r = 0.05$ ((a)-(f)) and $r = 0.25$ ((g)-(l)) respectively. Here, the ensemble size increases from 10 to 100 with an interval of 10. Accordingly, Figure 2 illustrates the performance curves of BAGGING, ADABOOST, ASSEMBLE and SEMIBOOST on the same data sets. These preliminary experiments indicate that, *in most cases*, the performance of all the algorithms does not significantly change within successive ensemble sizes and gradually levels out as the ensemble size grows to 60.

Therefore, for the sake of simplicity, three different ensemble sizes (i.e. m) are considered in this paper: a) $m = 20$ representing the case of *small-scale*

Table 4 Predictive accuracy of each comparing algorithm under *large-scale* ensemble size ($m = 100$). Bagging, AdaBoost, Assemble and SemiBoost are abbreviated as Bagg, AdaB, Assem and SemiB respectively.

	Algorithm									
	$r = 0.05$					$r = 0.25$				
Data Set	Udeed	Bagg	AdaB	Assem	SemiB	Udeed	Bagg	AdaB	Assem	SemiB
diabetes	0.653	0.653	0.649	0.650	**0.682**	0.700	0.692	**0.726**	0.694	0.696
heart	0.633	0.616	0.556	0.566	**0.664**	0.790	0.781	0.757	0.751	**0.792**
wdbc	0.661	0.638	0.714	0.663	**0.719**	0.852	0.805	**0.930**	0.916	0.825
austra	0.655	0.619	0.644	0.585	**0.701**	**0.824**	0.812	0.806	0.808	0.817
house	0.889	0.888	0.609	0.837	**0.890**	0.921	0.921	0.831	0.919	**0.924**
vote	**0.898**	**0.898**	0.667	0.889	0.895	0.930	0.930	0.902	0.926	**0.932**
vehicle	**0.813**	0.792	0.714	0.625	0.745	0.913	0.915	**0.930**	0.911	0.897
hepatitis	0.784	0.788	0.601	0.784	**0.794**	**0.797**	0.790	0.743	0.782	**0.797**
labor	**0.660**	0.650	0.500	0.554	0.610	**0.811**	0.808	0.683	0.756	0.809
ethn	0.871	0.867	**0.891**	0.826	0.782	**0.943**	0.942	0.938	0.939	0.932
ionosphere	**0.694**	0.663	0.669	0.662	0.684	0.780	0.721	**0.812**	0.779	0.747
kr_vs_kp	0.857	0.854	**0.880**	0.836	0.832	0.939	0.938	**0.945**	0.944	0.935
isolet	**0.963**	0.962	0.745	0.829	0.932	0.989	0.989	0.616	0.984	**0.990**
sonar	**0.575**	0.565	0.547	0.507	0.552	0.690	0.689	**0.713**	0.679	0.696
colic	0.729	**0.731**	0.663	0.674	0.690	0.784	**0.786**	0.741	0.745	0.763
credit_g	0.689	**0.696**	0.655	0.672	0.679	0.706	**0.711**	0.679	0.686	0.703
BCI	**0.517**	0.512	0.510	0.505	0.511	0.580	0.578	**0.620**	0.588	0.572
Digit1	**0.895**	0.888	0.845	0.850	0.892	0.940	0.940	0.927	0.925	**0.941**
COIL2	0.676	0.676	**0.742**	0.647	0.715	0.807	0.811	**0.870**	0.819	0.820
g241n	**0.665**	0.663	0.662	0.602	0.624	0.795	**0.796**	0.760	0.754	0.792
adult	0.836	**0.837**	0.817	0.830	N/A	**0.844**	**0.844**	0.840	0.843	N/A
web	0.974	0.974	0.971	**0.975**	N/A	**0.981**	0.980	0.980	**0.981**	N/A
ijcnn1	0.904	0.904	**0.919**	0.904	N/A	**0.906**	0.905	**0.906**	**0.906**	N/A
cod-rna	0.825	0.792	**0.938**	0.682	N/A	0.873	0.851	**0.945**	0.851	N/A
forest	0.707	0.703	**0.734**	0.688	N/A	0.705	0.703	**0.737**	0.698	N/A

ensemble; b) $m = 50$ representing the case of *medium-scale* ensemble; and c) $m = 100$ representing the case of *large-scale* ensemble.

Tables 2 to 4 report the detailed experimental results under *small-scale*, *medium-scale* and *large-scale* ensemble sizes respectively, when *few* ($r = 0.05$) or *abundant* ($r = 0.25$) labeled data is available. On each data set, the mean predictive accuracy out of 50 runs of each comparing algorithm is recorded, and the best performance is shown in boldface. In addition, SemiBoost fails to work on the five *large-scale* data sets, due to its demanding storage complexity, i.e. $\mathcal{O}((|\mathcal{L}| + |\mathcal{U}|)^2)$, to maintain the similarity matrix for labeled and unlabeled training examples.

For *small-scale* ensemble size (Table 2), when *few* labeled data is available ($r = 0.05$), Udeed ranks in *1st* place among the five comparing algorithms at 48% cases, in *2nd* place at 40% cases, in *3rd* place at 4% cases, and in *4th* or *5th* places at only 8% cases; When *abundant* labeled data is available ($r = 0.25$), Udeed ranks in *1st* place at 32% cases, in *2nd* place at 40% cases, in *3rd* place at 20% cases, and in *4th* or *5th* places at only 8% cases.

Table 5 Wilcoxon signed-ranks test (at 95% significance level) for UDEED versus each of the other comparing algorithms. The p-value out of the corresponding statistical test is shown in the brackets.

		UDEED **versus**			
Labeled Ratio	Ensemble Size	BAGGING	ADABOOST	ASSEMBLE	SEMIBOOST
$r = 0.05$	$m = 20$	**win** [*6.3e-3*]	**win** [*8.0e-3*]	**win** [*1.1e-4*]	**win** [*1.3e-2*]
	$m = 50$	**win** [*1.1e-2*]	**win** [*1.0e-2*]	**win** [*4.1e-5*]	**tie** [*5.9e-2*]
	$m = 100$	**win** [*3.2e-3*]	**win** [*3.8e-2*]	**win** [*4.0e-5*]	**tie** [*3.0e-1*]
$r = 0.25$	$m = 20$	**win** [*1.8e-2*]	**tie** [*1.5e-1*]	**win** [*6.9e-4*]	**win** [*2.6e-2*]
	$m = 50$	**win** [*7.6e-3*]	**tie** [*2.8e-1*]	**win** [*8.9e-3*]	**win** [*2.0e-2*]
	$m = 100$	**win** [*2.6e-2*]	**tie** [*5.3e-1*]	**win** [*1.1e-2*]	**win** [*4.0e-2*]

For *medium-scale* ensemble size (Table 3), when *few* labeled data is available ($r = 0.05$), UDEED ranks in *1st* place among the five comparing algorithms at 44% cases, in *2nd* place at 36% cases, in *3rd* place at 12% cases, and in *4th* or *5th* places at only 8% cases; When *abundant* labeled data is available ($r = 0.25$), UDEED ranks in *1st* place at 32% cases, in *2nd* place at 44% cases, in *3rd* place at 16% cases, and in *4th* or *5th* places at only 8% cases.

For *large-scale* ensemble size (Table 4), when *few* labeled data is available ($r = 0.05$), UDEED ranks in *1st* place among the five comparing algorithms at 36% cases, in *2nd* place at 52% cases, in *3rd* place at 8% cases, and in *4th* or *5th* places at only 4% cases; When *abundant* labeled data is available ($r = 0.25$), UDEED ranks in *1st* place at 28% cases, in *2nd* place at 48% cases, in *3rd* place at 20% cases, and in *4th* or *5th* places at only 4% cases.

To perform comparative analysis in a more well-founded way, we further examine the relative performance among the comparing algorithms based on statistical tests. Generally speaking, *Friedman test* may be the favorable choice (Demšar 2006) as several algorithms are compared over multiple data sets in this paper. Unfortunately, it is not directly applicable here as the ranks of SEMIBOOST on the five large-scale data sets are not available due to the missing results on them. As an alternative, we employ the Wilcoxon signed-ranks test (Demšar 2006; Wilcoxon 1945) to see whether UDEED is significantly different from each of the other comparing algorithms. Table 5 summarizes the statistical test results at 95% significance level, where the p-values of respective Wilcoxon signed-ranks tests are also reported in the brackets.

Results in Table 5 indicate that: a) UDEED outperforms BAGGING and ASSEMBLE under all the labeled ratios and ensemble sizes; b) UDEED achieves statistically comparable performance to ADABOOST when *abundant* ($r = 0.25$) labeled data is available, while outperforms ADABOOST when *few* ($r = 0.05$) labeled data is available; c) UDEED achieves statistically comparable performance to SEMIBOOST when *few* labeled data is available with *medium-scale* ($m = 50$) and *large-scale* ($m = 100$) ensemble sizes, while outperforms SEMIBOOST under other circumstances.

Table 6 Average ranks as well as critical difference (CD) for the post-hoc Nemenyi test (at 95% significance level) among UDEED, BAGGING, ADABOOST and ASSEMBLE. The lowest average rank is shown in boldface and a mark "⋆" is indicated if the average rank difference between UDEED and the comparing algorithm is larger than one CD.

		Average rank				
Labeled Ratio	Ensemble Size	UDEED	BAGGING	ADABOOST	ASSEMBLE	CD
$r = 0.05$	$m = 20$	**1.680**	2.360	2.640⋆	3.320⋆	0.938
	$m = 50$	**1.720**	2.240	2.840⋆	3.200⋆	0.938
	$m = 100$	**1.700**	2.240	2.880⋆	3.340⋆	0.938
$r = 0.25$	$m = 20$	**1.920**	2.500	2.600	2.980⋆	0.938
	$m = 50$	**1.900**	2.680	2.740	2.680	0.938
	$m = 100$	**1.960**	2.540	2.620	3.040⋆	0.938

Therefore, out of all the 24 statistical comparisons (2 labeled ratios × 3 ensemble sizes × 4 comparing algorithms), it is rather impressive that UDEED achieves significantly superior performance in 79.2% cases and no algorithms have once outperformed UDEED. The above results clearly validate that UDEED is highly competitive to other well-established ensemble learning methods, whenever *few* or *abundant* labeled data is available.

As a reference, we also performed Friedman test (together with post-hoc Nemenyi test (Demšar 2006) at 95% significance level) at the expense of excluding SEMIBOOST for comparison. A total of 18 statistical comparisons are conducted (2 labeled ratios × 3 ensemble sizes × 3 comparing algorithms), where the *average rank* of UDEED and each comparing algorithm as well as the *critical difference* (CD) are reported in Table 6. The results reveal that: a) UDEED achieves significantly superior performance in 44.4% cases, i.e. better than ADABOOST and ASSEMBLE with $r = 0.05$ under all ensemble sizes and better than ASSEMBLE with $r = 0.25$ under *small-scale* ($m = 20$) and *large-scale* ($m = 100$) ensemble sizes; b) No algorithms have once outperformed UDEED.

4.3 The Helpfulness of Unlabeled Data

As motivated in Section 1, 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 rather 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 7 reports the performance improvement (i.e. increase of predictive accuracy) of LCUD against LC and LCD under various settings. On each data set, the mean improved predictive accuracies out of 50 runs are recorded. Sim-

Table 7 Accuracy improvement for LcUd against Lc and Lcd under various labeled ratios and ensemble sizes.

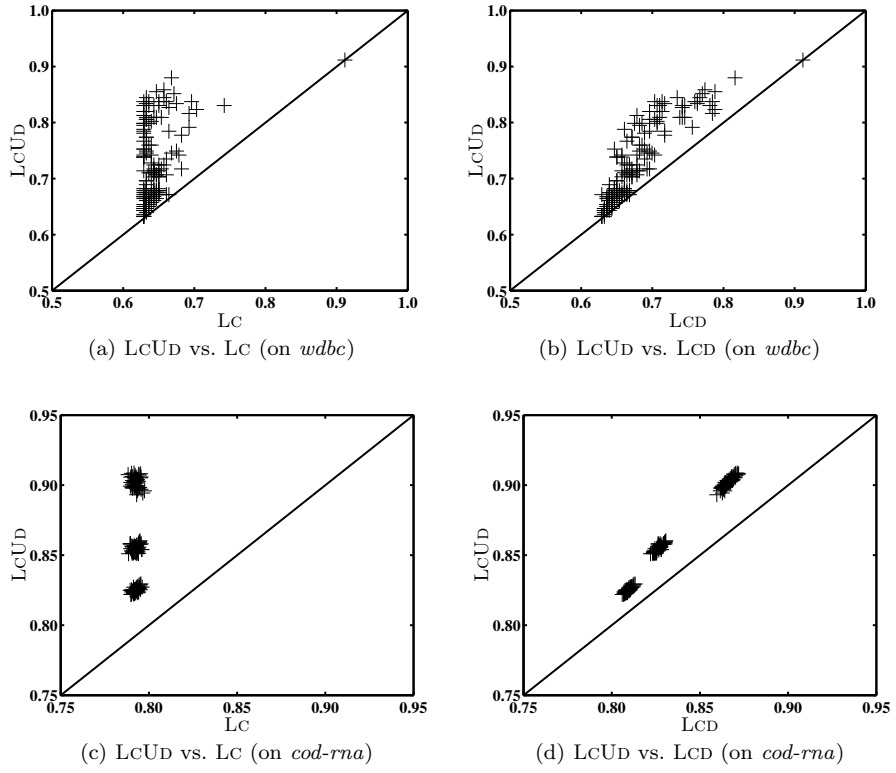
	LcUd **against**											
	Lc						Lcd					
	$r = 0.05$			$r = 0.25$			$r = 0.05$			$r = 0.25$		
Data Set	m=20	m=50	m=100	m=20	m=50	m=100	m=20	m=50	m=100	m=20	m=50	m=100
diabetes	0.017	0.004	0.001	0.034	0.019	0.008	0.010	0.002	0.001	0.011	0.009	0.004
heart	0.049	0.029	0.023	0.023	0.009	0.006	0.023	0.013	0.010	0.009	0.003	0.004
wdbc	0.161	0.059	0.021	0.127	0.075	0.047	0.090	0.038	0.013	0.033	0.031	0.023
austra	0.089	0.067	0.036	0.022	0.015	0.010	0.026	0.032	0.018	0.004	0.006	0.005
house	-0.004	0.001	0.001	0.003	-0.001	0.001	-0.001	-0.001	-0.001	0.002	0.000	0.001
vote	0.001	0.001	-0.001	0.002	0.001	0.001	0.001	0.001	-0.001	0.001	0.001	0.001
vehicle	0.029	0.017	0.010	0.005	0.002	0.001	0.007	0.007	0.006	0.003	0.001	0.001
hepatitis	-0.018	-0.004	-0.005	0.010	0.005	0.008	-0.007	-0.002	-0.004	0.003	0.001	0.005
labor	0.010	0.013	0.010	0.003	0.004	0.004	-0.002	0.006	0.007	-0.007	0.007	0.004
ethn	0.011	0.005	0.003	0.002	0.001	0.001	0.006	0.004	0.003	0.001	0.001	0.001
ionosphere	0.028	0.038	0.031	0.073	0.076	0.057	-0.003	0.011	0.020	0.015	0.022	0.029
kr_vs_kp	0.005	0.004	0.002	0.002	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001
isolet	0.007	0.007	0.003	0.001	0.001	0.001	0.004	0.006	0.003	0.001	0.001	0.001
sonar	0.013	0.005	0.005	0.001	0.003	0.001	0.006	0.002	0.003	0.002	-0.001	0.001
colic	-0.010	-0.004	-0.001	-0.006	-0.003	-0.001	-0.005	-0.002	0.001	-0.003	-0.003	0.001
credit_g	-0.039	-0.018	-0.007	-0.019	-0.008	-0.005	-0.021	-0.010	-0.003	-0.009	-0.004	-0.002
BCI	0.008	-0.001	0.001	0.006	0.003	0.002	0.005	0.001	-0.001	0.005	0.002	0.002
Digit1	0.006	0.007	0.005	0.001	0.001	0.001	0.003	0.003	0.003	0.001	0.001	0.001
COIL2	0.012	0.002	0.002	-0.001	-0.004	-0.003	0.007	0.002	0.002	0.001	-0.001	-0.002
g241n	0.002	0.003	0.002	0.001	0.001	-0.001	0.001	0.003	0.002	-0.001	0.001	-0.001
adult	-0.005	-0.002	-0.001	-0.009	-0.002	-0.001	-0.003	-0.001	-0.001	-0.006	-0.002	-0.001
web	0.001	0.001	0.001	0.001	0.001	0.000	0.001	0.001	0.001	0.001	0.001	0.000
ijcnn1	-0.001	0.000	0.000	0.008	0.001	0.001	-0.001	0.000	0.000	0.006	0.001	0.001
cod-rna	0.109	0.062	0.033	0.069	0.041	0.023	0.036	0.029	0.016	0.022	0.018	0.011
forest	0.003	0.002	0.001	0.003	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 8 Wilcoxon signed-ranks test (at 95% significance level) for LcUd versus Lc and Lcd. The p-value out of the corresponding statistical test is shown in the brackets.

		LcUd **versus**	
Labeled Ratio	Ensemble Size	Lc	Lcd
$r = 0.05$	$m = 20$	**win** [*1.0e-2*]	**win** [*4.2e-2*]
	$m = 50$	**win** [*3.7e-3*]	**win** [*3.1e-3*]
	$m = 100$	**win** [*2.7e-3*]	**win** [*4.4e-3*]
$r = 0.25$	$m = 20$	**win** [*3.4e-3*]	**win** [*1.8e-2*]
	$m = 50$	**win** [*8.7e-3*]	**win** [*1.1e-2*]
	$m = 100$	**win** [*4.7e-3*]	**win** [*2.0e-3*]

ilar as Table 5, Wilcoxon signed-ranks test is employed to see whether LcUd is significantly different from its two counterparts. Table 8 summarizes the statistical test results at 95% significance level, where the p-values of respective Wilcoxon signed-ranks tests are reported in the brackets.

Results in Table 8 indicate that, out of all the 12 statistical comparisons (2 labeled ratios × 3 ensemble sizes × 2 comparing algorithms), it is rather impressive that LcUd achieves significantly superior performance in all 100% cases. The above results clearly validate that, by exploiting unlabeled data in the specific way of helping augment ensemble diversity, Udeed (LcUd) is



(a) LCUD vs. LC (on *wdbc*) (b) LCUD vs. LCD (on *wdbc*)

(c) LCUD vs. LC (on *cod-rna*) (d) LCUD vs. LCD (on *cod-rna*)

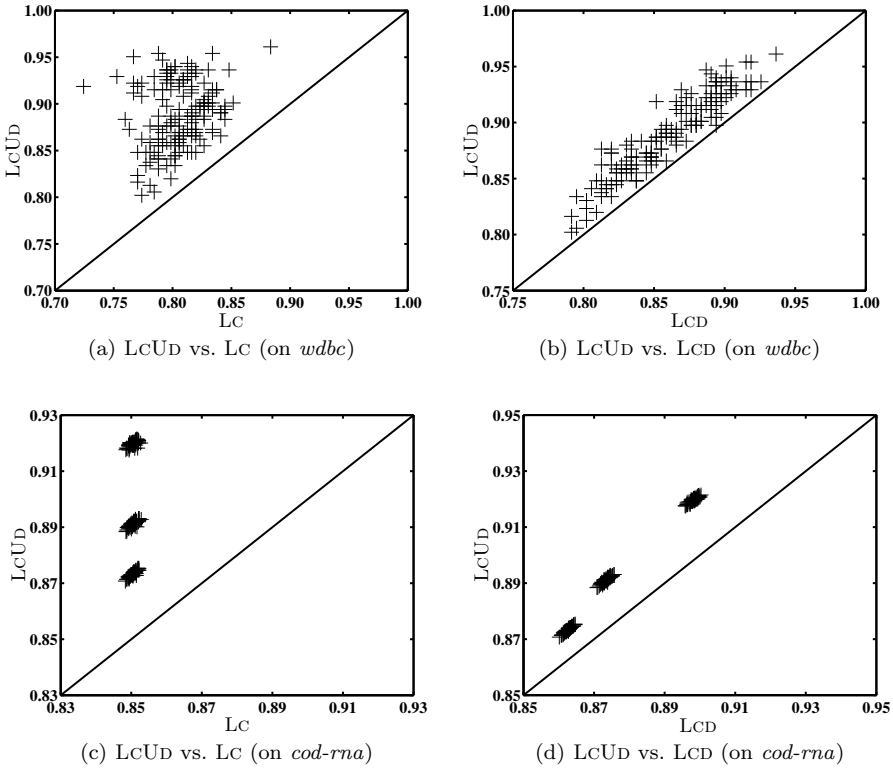
Fig. 3 Illustrative scatter plots between LCUD and LC, LCD in terms of predictive accuracy with $r = 0.05$. *Left column*: LCUD vs. LC, *Right column*: LCUD vs. LCD.

capable of achieving better performance than its counterparts (LC and LCD) which do not consider employing unlabeled in ensemble generation, whenever *few* or *abundant* labeled data is available.

Furthermore, for more intuitionistic illustration, Figure 3 gives the scatter plots between LCUD and its compared implementations LC and LCD on one regular-scale data set (*wdbc*) and one large-scale data set (*cod-rna*), with $r = 0.05$. Accordingly, Figure 4 gives the scatter plots with $r = 0.25$. In each plot, when the ensemble size is fixed, the predictive accuracy of LCUD in each of the 50 runs is plotted against the compared implementation with a marker '+' in the figure. Obviously, LCUD achieves better performance than LC and LCD as the majority of markers lie above the diagonal.

4.4 Diversity Analysis

To clearly verify that UDEED (LCUD) does increase the diversity among base learners after generating the ensemble by utilizing unlabeled data, additional



(a) LCUD vs. LC (on *wdbc*) (b) LCUD vs. LCD (on *wdbc*)

(c) LCUD vs. LC (on *cod-rna*) (d) LCUD vs. LCD (on *cod-rna*)

Fig. 4 Illustrative scatter plots between LCUD and LC, LCD in terms of predictive accuracy with $r = 0.25$. *Left column*: LCUD vs. LC, *Right column*: LCUD vs. LCD.

experiments are analyzed in this subsection based on several existing diversity measures. Specifically, four diversity measures summarized in (Kuncheva and Whitaker 2003) are considered, whose values are calculated based on the *oracle* outputs of base learners, i.e. correct or incorrect prediction for the class label.

Let m denote the number of base classifiers in the ensemble and N denote 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 ($1 \leq i \leq m$, $1 \leq j \leq N$). Otherwise, $o_{ij} = 0$. The formal definitions of the four diversity measures are as follows:

• *Disagreement measure* (DIS) (Skalak 1996):

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

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

Here, dis_{ik} represents the ratio between the number of examples on which a pair of base classifiers make opposite predictions to the total number of examples in $\mathcal{T}$.

• *Double-fault measure* (DF) (Giacinto and Roli 2001):

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

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

Here, df_{ik} represents the proportion of examples in $\mathcal{T}$ which have been misclassified by both base classifiers.

• *Entropy measure* (ENT) (Cunningham and Carney 2000):

$$\mathrm{ENT} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{m - \lceil m/2 \rceil} \min \{L_j, m - L_j\}, \quad \text{where}$$

$$L_j = \sum_{i=1}^{m} o_{ij}$$

Here, L_j represents the number of base classifiers which make correct predictions for the j-th test example.

• *Coincident failure diversity* (CFD) (Partridge and Krzanowski 1997):

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

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

Here, $1_{[\![\pi]\!]}$ is an indicator function which takes value of 1 if predicate π holds and 0 otherwise. Accordingly, p_i represents the probability that i out of m base classifiers give incorrect predictions on a randomly drawn example from $\mathcal{T}$.

Note that DIS and DF are *pairwise* measures which evaluate ensemble diversity by leveraging the relations between each pair of base classifiers, while ENT and CFD are *non-pairwise* measures which evaluate ensemble diversity from a more holistic viewpoint. In this paper, 1-DF is used instead of DF such

Table 9 Wilcoxon signed-ranks test (at 95% significance level) for UDEED's FINAL ensemble versus its INITIAL ensemble in terms of four diversity measures. The p-value out of the corresponding statistical test is shown in the brackets.

		FINAL ensemble **versus** INITIAL ensemble			
Labeled Ratio	Ensemble Size	DIS	DF	ENT	CFD
$r = 0.05$	$m = 20$	**win** [*4.3e-2*]	**tie** [*8.8e-1*]	**win** [*3.0e-2*]	**tie** [*1.8e-1*]
	$m = 50$	**win** [*3.2e-2*]	**tie** [*8.3e-2*]	**win** [*4.6e-2*]	**tie** [*7.2e-2*]
	$m = 100$	**win** [*4.0e-2*]	**win** [*8.6e-3*]	**win** [*4.0e-2*]	**win** [*3.2e-2*]
$r = 0.25$	$m = 20$	**win** [*2.6e-2*]	**tie** [*6.8e-2*]	**win** [*2.8e-2*]	**win** [*4.7e-3*]
	$m = 50$	**win** [*3.2e-2*]	**win** [*3.7e-2*]	**win** [*3.0e-2*]	**win** [*7.4e-3*]
	$m = 100$	**win** [*2.8e-2*]	**win** [*1.4e-2*]	**win** [*3.2e-2*]	**win** [*6.6e-3*]

Table 10 Wilcoxon signed-ranks test (at 95% significance level) for UDEED versus BAGGING in terms of four diversity measures. The p-value out of the corresponding statistical test is shown in the brackets.

		UDEED **versus** BAGGING			
Labeled Ratio	Ensemble Size	DIS	DF	ENT	CFD
$r = 0.05$	$m = 20$	**win** [*4.3e-2*]	**tie** [*1.7e-1*]	**win** [*4.0e-2*]	**win** [*4.9e-2*]
	$m = 50$	**win** [*4.6e-2*]	**win** [*4.5e-2*]	**tie** [*5.2e-2*]	**win** [*6.1e-3*]
	$m = 100$	**win** [*4.3e-2*]	**win** [*2.1e-2*]	**win** [*4.3e-2*]	**tie** [*1.4e-1*]
$r = 0.25$	$m = 20$	**win** [*9.3e-3*]	**tie** [*9.8e-2*]	**win** [*1.2e-2*]	**win** [*1.2e-2*]
	$m = 50$	**tie** [*7.8e-2*]	**win** [*1.8e-2*]	**tie** [*9.3e-2*]	**win** [*2.8e-2*]
	$m = 100$	**win** [*2.1e-2*]	**win** [*1.0e-2*]	**win** [*2.6e-2*]	**win** [*1.2e-2*]

Table 11 Wilcoxon signed-ranks test (at 95% significance level) for UDEED versus ADABOOST in terms of four diversity measures. The p-value out of the corresponding statistical test is shown in the brackets.

		UDEED **versus** ADABOOST			
Labeled Ratio	Ensemble Size	DIS	DF	ENT	CFD
$r = 0.05$	$m = 20$	**loss** [*3.0e-3*]	**tie** [*6.2e-2*]	**loss** [*5.1e-3*]	**tie** [*1.2e-1*]
	$m = 50$	**loss** [*6.1e-3*]	**tie** [*5.1e-2*]	**loss** [*1.4e-2*]	**win** [*4.9e-2*]
	$m = 100$	**loss** [*1.4e-3*]	**win** [*2.8e-2*]	**loss** [*2.4e-3*]	**win** [*4.5e-2*]
$r = 0.25$	$m = 20$	**loss** [*1.8e-5*]	**tie** [*8.8e-1*]	**loss** [*2.4e-5*]	**loss** [*4.3e-3*]
	$m = 50$	**loss** [*1.8e-5*]	**tie** [*6.2e-1*]	**loss** [*1.8e-5*]	**tie** [*7.2e-2*]
	$m = 100$	**loss** [*2.7e-5*]	**tie** [*5.1e-1*]	**loss** [*2.7e-5*]	**tie** [*2.1e-1*]

that for all the measures, the *greater* the value the *higher* the diversity. In addition, all the four measures vary between 0 and 1.

Table 9 compares UDEED's *initial* diversity after ensemble initialization with its *final* diversity after ensemble learning under various settings. Similar as Table 5, Wilcoxon signed-ranks test is employed to see whether the diversity

of UDEED's final ensemble is significantly different from its initial ensemble. Statistical test results at 95% significance level are summarized in Table 8, where the p-values of respective Wilcoxon signed-ranks tests are reported in the brackets.

Results in Table 9 indicate that: a) UDEED significantly increases the initial ensemble diversity in terms of DIS and ENT under all the labeled ratios and ensemble sizes; b) UDEED significantly increases the initial ensemble diversity in terms of DF when *few* ($r = 0.05$) labeled data is available with *large-scale* ($m = 100$) ensemble size, and when *abundant* ($r = 0.25$) labeled data is available with *medium-scale* ($m = 50$) and *large-scale* ($m = 100$) ensemble sizes; c) UDEED significantly increases the initial ensemble diversity in terms of CFD when *few* labeled data is available with *large-scale* ensemble size, and when *abundant* labeled data is available.

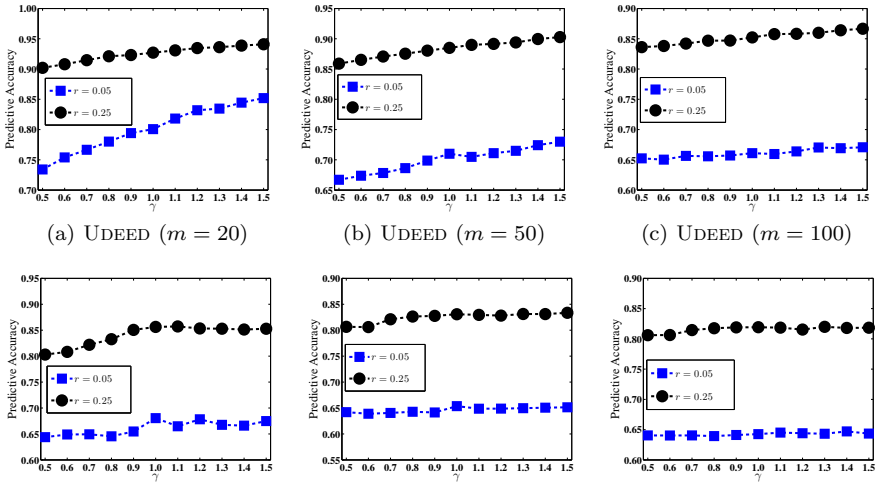
Therefore, out of all the 24 statistical comparisons (2 labeled ratios $\times$ 3 ensemble sizes $\times$ 4 diversity measures), it is rather impressive that UDEED significantly increases the initial ensemble diversity in 79.2% cases and never significantly decreases the initial ensemble diversity. The above results clearly validate that UDEED can effectively exploit unlabeled data to help augment ensemble diversity, whenever *few* or *abundant* labeled data is available.

Similar as Table 9, Table 10 and 11 also compare the final ensemble diversity produced by UDEED to those of BAGGING and ADABOOST respectively, where the latter two algorithms both work under supervised setting without considering unlabeled data for diversity augmentation. Table 10 indicates that, out of all the 24 statistical comparisons, UDEED attains significantly higher ensemble diversity than BAGGING in 75% cases and never produces significantly lower ensemble diversity.

However, as indicated in Table 11, although ADABOOST attains significantly higher ensemble diversity than UDEED in 54.2% cases, this may be achieved at the expense of base learners' accuracies. To verify this, Wilcoxon signed-ranks test is again employed to compare the mean accuracy of the base learners in the ensemble returned by UDEED and ADABOOST. Statistical test results (at 95% significance level) reveal that, when *few* labeled data is available, UDEED achieves significantly superior mean base accuracy than ADABOOST with *small-scale* ensemble size ($p < 6.7e\text{-}4$), *medium-scale* ensemble size ($p < 9.0e\text{-}4$), and *large-scale* ensemble size ($p < 1.9e\text{-}4$); When *abundant* labeled data is available, UDEED achieves significantly superior mean base accuracy than ADABOOST with *small-scale* ensemble size ($p < 1.6e\text{-}5$), *medium-scale* ensemble size ($p < 1.8e\text{-}5$), and *large-scale* ensemble size ($p < 3.5e\text{-}4$). These results show that UDEED is capable of maintaining a healthy equilibrium between the ensemble diversity and accuracies.

5 Discussion

In this section, several issues related to the UDEED approach are further discussed. Firstly, the algorithmic behaviors of UDEED are analyzed regarding the



(a) UDEED ($m = 20$) (b) UDEED ($m = 50$) (c) UDEED ($m = 100$)

(d) UDEED-NCL ($m = 20$) (e) UDEED-NCL ($m = 50$) (f) UDEED-NCL ($m = 100$)

Fig. 5 Performance of UDEED (first row) and UDEED-NCL (second row) on the *wdbc* data set changes as the cost parameter γ increases under various settings.

optimization of its objective function. After that, the relationships between UDEED and two well-established techniques for ensemble diversity augmentation commonly used in supervised setting, i.e. negative correlation learning (Liu and Yao 1999a,b) and artificial examples construction (Melville and Mooney 2003; Melville 2005), are discussed.

5.1 Algorithmic Behavior

In this paper, the cost parameter γ (Eq.(1)) controlling the tradeoff between the empirical and diversity losses of UDEED is set to the default value of 1. Figure 5 (first row) illustrates the impact of γ on UDEED's performance on one representative data set *wdbc*, where γ increases from 0.5 to 1.5 with an interval of 0.1. It is obvious that in most cases, increasing the value of γ wouldn't jeopardize the performance of UDEED. To avoid over-emphasizing the diversity loss on the objective function, we choose to set $\gamma = 1$ in the experiments reported in Section 4. However, better performance could be expected if certain strategies such as cross-validation are utilized to finely tune the cost parameter.

As shown in Subsection 3.2, the target model of UDEED is found by optimizing Eq.(1) based on *gradient decent*-based techniques. Therefore, the returned model would correspond to a *local* optimal solution. To roughly illustrate the possibility of getting stuck in a "bad" local solution, Table 12 summarizes the statistics of accuracy out of 50 runs, where the minimal, maximal, and mean (together with standard deviation) values are reported for the *wdbc* and *cod-rna* data sets. Intuitively, values that fall one standard deviation

Table 12 Accuracy statistics out of 50 runs on the *wdbc* and *cod-rna* data sets under various settings, where values falling one standard deviation below the mean value are regarded as "bad" local solutions.

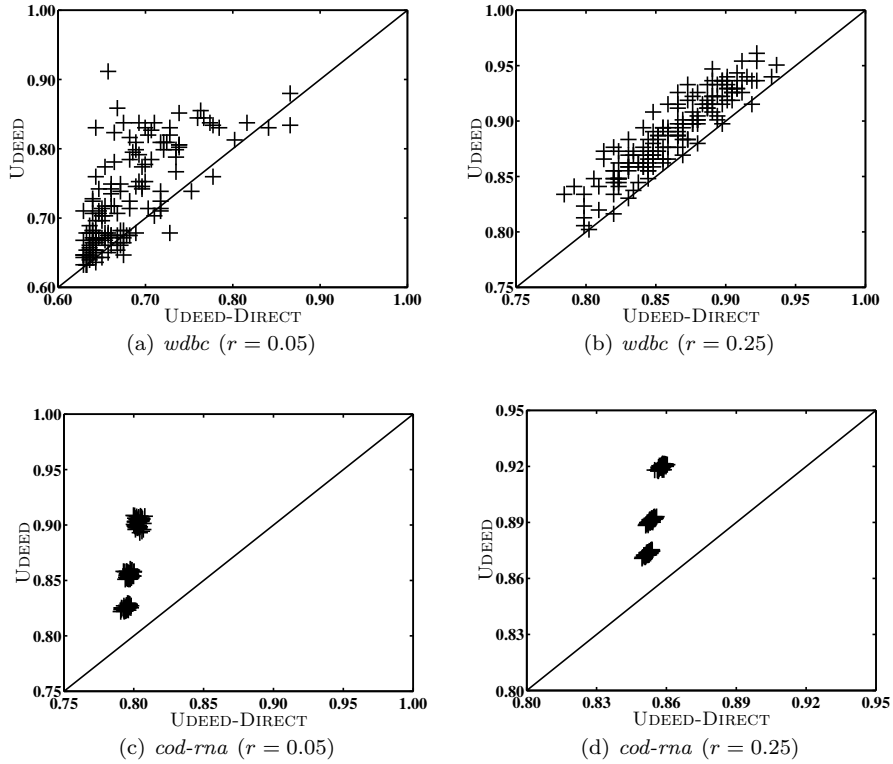
	wdbc					
	$r = 0.05$			$r = 0.25$		
	$m = 20$	$m = 50$	$m = 100$	$m = 20$	$m = 50$	$m = 100$
MIN	0.714	0.643	0.633	0.883	0.848	0.802
MAX	0.880	0.912	0.717	0.961	0.915	0.894
MEAN±STD. DEVIATION	0.802	0.710	0.661	0.927	0.885	0.852
	±0.039	±0.051	±0.020	±0.014	±0.017	±0.021
PROB. OF "BAD" LOC. SOL.	**22%**	**20%**	**12%**	**14%**	**16%**	**14%**

	cod-rna					
	$r = 0.05$			$r = 0.25$		
	$m = 20$	$m = 50$	$m = 100$	$m = 20$	$m = 50$	$m = 100$
MIN	0.893	0.851	0.822	0.918	0.888	0.871
MAX	0.909	0.860	0.829	0.922	0.893	0.875
MEAN±STD. DEVIATION	0.902	0.855	0.825	0.920	0.891	0.873
	±0.004	±0.003	±0.002	±0.001	±0.001	±0.001
PROB. OF "BAD" LOC. SOL.	**16%**	**18%**	**16%**	**16%**	**16%**	**14%**

below the mean value will be regarded as "bad" local solutions. Accordingly, the probability of obtaining a "bad" local solution can be calculated as shown in Table 12. It is revealed that UDEED will get stuck in a "bad" local solution with no more than 25% probability under either data set and various settings.

As shown in Eq.(1), UDEED's objective function $V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$ is composed of two different terms, i.e. the *empirical loss* term $V_{emp}(\boldsymbol{f}, \mathcal{L})$ and the *diversity loss* term $V_{div}(\boldsymbol{f}, \mathcal{D})$. To show whether these two terms are optimized with comparable scales, we calculated the relative differences between the two losses, i.e. $\Delta_{loss} = |V_{emp} - V_{div}|/V$ ($\gamma = 1$), on the five large-scale data sets. If the two loss terms are on quite different scales, the value of Δ_{loss} would be *rather close* to 1. Out of all the 1500 calculations (2 labeled ratios × 3 ensemble sizes × 5 large-scale data sets × 50 runs), Δ_{loss} is less than 0.3 in 28.9% cases, less than 0.5 in 71% cases, and less than 0.8 in 100% cases. These results indicate that in most cases, UDEED could maintain a reasonable balance between the empirical loss term and diversity loss term with γ fixed to 1.

In addition to the above algorithmic behaviors, another implementation issue regarding UDEED (i.e. LCUD) lies in its specific gradient descent strategy. As shown in Subsection 3.2, the objective function of UDEED is minimized by using the intermediate model returned by LCD as the *starting point* for gradient descent optimization. To show the importance of the above choice, another version of UDEED named UDEED-DIRECT is implemented which directly invokes the minimization procedure without exploiting intermediate model.



(a) *wdbc* ($r = 0.05$) (b) *wdbc* ($r = 0.25$)

(c) *cod-rna* ($r = 0.05$) (d) *cod-rna* ($r = 0.25$)

Fig. 6 Scatter plots between UDEED and UDEED-DIRECT on the *wdbc* and *cod-rna* data sets in terms of predictive accuracy. *Left column*: $r = 0.05$, *Right column*: $r = 0.25$.

Figure 6 gives the scatter plots between UDEED and UDEED-DIRECT on the *wdbc* and *cod-rna* data sets. In each plot, when the ensemble size is fixed, the predictive accuracy of UDEED in each of the 50 runs is plotted against UDEED-DIRECT with a marker '+' in the figure. Obviously, UDEED achieves better performance than UDEED-DIRECT as the majority of markers lie above the diagonal. This clearly validates that employing the intermediate model returned by LCD as the starting point is quite crucial for the success of UDEED.

5.2 Diversity Augmentation

UDEED aims to build strong ensembles by exploiting unlabeled data for diversity augmentation. On the other hand, the idea of enhancing ensemble diversity has been widely investigated in ensemble learning under supervised setting. In this subsection, two existing strategies closely related to UDEED will be discussed in more details.

As shown in Eq.(3), UDEED measures ensemble diversity by considering the *predictive difference* between each pair of base classifiers on a specified data set $\mathcal{D}$. Here, the predictive difference could be viewed as a specific quantification of *correlation* between two classifiers, and maximizing diversity on $\mathcal{D}$ is effectively equivalent to *punish* correlations among base learners in the ensemble. Actually, the idea of *decorrelation* has been investigated in learning neural network ensembles (Rosen 1996), and later developed into the well-known techniques named *negative correlation learning* (NCL) (Brown and Wyatt 2003; Chen and Yao 2009; Liu and Yao 1999a,b; McKay and Abbass 2001). NCL aims to train an ensemble by introducing a correlation penalty term to the cost function of each individual learner.

NCL was proposed for ensemble learning under supervised setting. Here we generalize it to unlabeled data and implement the *diversity loss* term in Eq.(1) by enforcing negative correlation constraints (Liu and Yao 1999a,b) on the unlabeled data as follows:

$$\begin{aligned} V_{div}(\boldsymbol{f},\mathcal{D}) &= V_{div}(\boldsymbol{f},\mathcal{U}) \\ &= \frac{1}{m}\sum_{k=1}^{m} p_k \\ &= \frac{1}{m}\sum_{k=1}^{m}\cdot\frac{1}{U}\sum_{i=L+1}^{L+U}(f_k(\boldsymbol{x}_i)-\boldsymbol{f}(\boldsymbol{x}_i))\sum_{k'\neq k}(f_{k'}(\boldsymbol{x}_i)-\boldsymbol{f}(\boldsymbol{x}_i)) \\ &= -\frac{1}{mU}\sum_{k=1}^{m}\sum_{i=L+1}^{L+U}(f_k(\boldsymbol{x}_i)-\boldsymbol{f}(\boldsymbol{x}_i))^2 \end{aligned} \tag{8}$$

Where $\mathcal{D}$ corresponds to the unlabeled data set $\mathcal{U}$, p_k represents the diversity loss of the k-th classifier f_k on $\mathcal{D}$, and $\boldsymbol{f}(\boldsymbol{x}_i) = \frac{1}{m}\sum_{k=1}^{m} f_k(\boldsymbol{x}_i)$ is the ensemble output on $\boldsymbol{x}_i$. The resulting variant of UDEED is named as UDEED-NCL. Here, logistic regression is again utilized as the base learner and the target model $\boldsymbol{f}^*$ is also found by employing gradient descent-based techniques.[5]

It is known that using an ensemble to generate artificial data, and then using these data to train another learner is beneficial when the data sample does not capture the whole data distribution or contains noise (Zhou and Jiang 2004). Melville and Mooney(2003; 2005) argued that by using artificial example construction, the diversity can be augmented. They proposed the DECORATE approach which builds an ensemble in an iterative manner. In each iteration, the labeled training set is enlarged with a number of randomly constructed artificial examples, which are given labels that *disagree* with the decisions of current ensemble so as to encourage diversity when building the new ensemble base classifier. In this paper, we re-implemented the DECORATE

[5] Similar to UDEED, Figure 5 (second row) also illustrates the impact of γ on UDEED-NCL's performance on one representative data set *wdbc*, where γ increases from 0.5 to 1.5 with an interval of 0.1. It is obvious that in most cases, increasing the value of γ wouldn't jeopardize the performance of UDEED-NCL. Furthermore, UDEED-NCL tends to perform stably when γ increases to 1. Therefore, we choose to set $\gamma = 1$ for UDEED-NCL in the following experiments.

Table 13 Accuracy difference for UDEED against UDEED-NCL and DECORATE under various settings.

	UDEED **against**											
	UDEED-NCL						DECORATE					
	$r = 0.05$			$r = 0.25$			$r = 0.05$			$r = 0.25$		
Data Set	m=20	m=50	m=100	m=20	m=50	m=100	m=20	m=50	m=100	m=20	m=50	m=100
diabetes	0.015	0.004	0.001	0.026	0.014	0.006	-0.013	-0.038	-0.043	-0.016	-0.011	-0.043
heart	0.045	0.041	0.022	0.005	0.006	0.006	0.033	0.027	0.037	0.021	0.014	0.000
wdbc	0.128	0.058	0.018	0.074	0.056	0.034	-0.014	-0.010	-0.016	0.001	-0.001	-0.008
austra	0.060	0.044	0.020	0.011	0.009	0.009	0.001	0.003	0.007	0.002	0.009	0.012
house	0.001	0.001	-0.001	0.001	0.001	0.001	0.009	0.005	0.004	0.008	0.009	0.008
vote	0.001	0.001	0.001	0.001	0.001	0.001	0.013	0.013	0.012	0.007	0.006	0.007
vehicle	0.010	0.006	0.006	0.001	0.001	-0.002	0.013	0.001	0.001	0.005	0.001	0.016
hepatitis	-0.023	-0.008	-0.002	0.008	0.003	0.007	0.064	0.056	0.045	0.034	0.007	0.007
labor	0.010	0.013	0.010	-0.007	0.001	0.001	0.040	0.012	0.002	0.005	0.003	0.008
ethn	0.001	0.002	0.001	0.001	0.001	0.001	0.003	0.008	0.007	0.002	0.003	0.001
ionosphere	0.006	0.028	0.029	0.027	0.042	0.037	-0.006	-0.011	-0.012	-0.007	-0.018	-0.003
kr_vs_kp	0.005	0.002	0.002	0.002	0.001	0.001	0.004	0.010	0.014	0.006	0.008	0.010
isolet	0.002	0.000	-0.001	0.001	0.001	0.001	0.050	0.025	0.022	0.024	0.003	0.004
sonar	0.016	0.003	0.004	0.007	-0.004	0.001	0.005	-0.002	-0.007	0.001	-0.002	-0.001
colic	-0.013	-0.005	-0.003	-0.004	-0.001	-0.001	0.052	0.046	0.049	0.018	0.022	0.028
credit_g	-0.033	-0.015	-0.004	-0.016	-0.006	-0.003	-0.006	0.021	0.048	-0.006	0.005	0.010
BCI	-0.001	0.004	0.003	0.004	0.001	0.004	0.001	0.003	0.004	0.004	0.008	0.009
Digit1	0.012	0.004	0.004	-0.001	0.001	-0.001	0.016	0.005	0.001	0.002	0.003	0.001
COIL2	0.014	-0.002	0.002	-0.001	-0.002	-0.004	-0.013	-0.021	-0.018	-0.012	-0.011	-0.012
g241n	0.004	-0.001	0.001	0.000	0.001	-0.002	0.033	0.026	0.014	0.001	0.002	0.003
adult	-0.005	-0.001	-0.001	-0.009	-0.002	-0.001	-0.006	-0.002	-0.001	-0.009	-0.002	-0.001
web	0.001	0.003	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
ijcnn1	-0.001	-0.001	-0.001	0.008	0.001	0.001	-0.001	-0.001	-0.001	0.006	0.001	0.002
cod-rna	0.009	0.002	0.003	0.010	0.001	0.003	0.076	0.029	0.002	0.069	0.040	0.022
forest	0.004	0.002	0.004	0.003	0.002	0.002	0.007	0.009	0.006	0.004	0.001	0.001

Table 14 Wilcoxon signed-ranks test (at 95% significance level) for UDEED versus UDEED-NCL and DECORATE. The p-value out of the corresponding statistical test is shown in the brackets.

		UDEED **versus**	
Labeled Ratio	Ensemble Size	UDEED-NCL	DECORATE
$r = 0.05$	$m = 20$	**win** [*2.3e-2*]	**win** [*1.9e-2*]
	$m = 50$	**win** [*2.3e-2*]	**win** [*2.3e-2*]
	$m = 100$	**win** [*8.0e-3*]	**tie** [*5.8e-2*]
$r = 0.25$	$m = 20$	**win** [*2.8e-2*]	**win** [*4.0e-2*]
	$m = 50$	**win** [*2.1e-2*]	**win** [*3.0e-2*]
	$m = 100$	**win** [*2.8e-2*]	**win** [*2.2e-2*]

approach for comparative studies. For the sake of fair comparison, the number of artificial examples constructed by DECORATE in each iteration is the same as the number of unlabeled data used by UDEED, i.e. $|\mathcal{U}|$.

Table 13 reports the difference in performance of UDEED against UDEED-NCL and DECORATE under various settings. On each data set, the mean predictive accuracy differences out of 50 runs are recorded. Wilcoxon signed-ranks test is employed to see whether UDEED is significantly different from UDEED-NCL and DECORATE. Table 14 summarizes the statistical test results at 95%

significance level, where the p-values of respective Wilcoxon signed-ranks tests are reported in the brackets.

As shown in Table 14, UDEED achieves significantly superior performance than UDDED-NCL under different labeled ratios and ensemble sizes. Note that for UDEED, the diversity is calculated based on the pairwise predictive difference between each pair of base classifiers where no ground-truth labels on the unlabeled data are assumed (Eq.(3)). While for UDEED-NCL, the diversity is calculated based on the diversity loss of each individual base classifier where the outputs of ensemble implicitly serve as the ground-truth labels on the unlabeled data (Eq.(8)).

In addition, Table 14 shows that UDEED achieves statistically comparable performance to DECORATE under one setting ($r = 0.05$, $m = 100$), and achieves significantly superior performance under all the other settings. Note that the working mechanisms of using unlabeled data (UDEED) and using artificial data (DECORATE) are quite different. By using unlabeled data, the learning system can exploit the underlying distributional information concealed in unlabeled data. While by using artificial data, the randomly constructed artificial examples may not truly reflect the underlying distributions and therefore lead to possible overfitting of the learned system.

6 Conclusion

There have been many works trying to maximize diversity for ensemble construction, yet they are mainly based on using labeled data. On the other hand, there were some studies trying to use unlabeled data, yet they focus on using unlabeled data to improve accuracy. This paper proposes a new approach on ensemble learning with unlabeled data (Zhang and Zhou 2010), which works by maximizing accuracy on labeled data while maximizing diversity on unlabeled data. The major contribution of our work is to use unlabeled data to augment diversity, which suggests a new direction for ensemble design.

Extensive experiments on twenty-five data sets 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 and extending to categorical features and multi-class classification is a direction very worth studying.

Acknowledgements

We want to thank the action editor and the anonymous reviewers for their helpful comments and suggestions.

References

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

Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, pp 92–100

Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140

Brown G, Wyatt JL (2003) The use of the ambiguity decomposition in neural network ensemble learning methods. In: Proceedings of the 20th International Conference on Machine Learning, Washington D.C., pp 67–74

d'Alché Buc F, Grandvalet Y, Ambroise C (2002) Semi-supervised marginboost. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, pp 553–560

Chapelle O, Schölkopf B, Zien A (2006) Semi-Supervised Learning. MIT Press, Cambridge, MA

Chen HH, Yao X (2009) Regularized negative correlation learning for neural network ensemble. IEEE Transactions on Neural Networks 20(12):1962–1979

Chen K, Wang S (2008) Regularized boost for semi-supervised learning. In: Platt JC, Koller D, Singer Y, Roweis S (eds) Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, MA, pp 281–288

Chen K, Wang S (2011) Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1):129–143

Cunningham P, Carney J (2000) Diversity versus quality in classification ensembles based on feature selection. Tech. Rep. TCD-CS-2000-02, Department of Computer Science, Trinity College Dublin, Dublin, Ireland

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7(Jan):1–30

Dietterich TG (2000) Ensemble methods in machine learning. In: Proceedings of the 1st International Workshop on Multiple Classifier Systems, Cagliari, Italy, pp 1–15

Efron B, Tibshirani R (1993) An Introduction to the Boostrap. Chapman & Hall, New York

Frank A, Asuncion A (2010) UCI machine learning repository [http://archive.ics.uci.edu/ml]. Tech. rep., School of Information and Computer Science, University of California, Irvine, CA

Freund Y, Schapire RE (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi PMB (ed) Lecture Notes in Computer Science 904, Springer, Berlin, pp 23–37

Giacinto G, Roli F (2001) Design of effective neural network ensembles for image classification processes. Image and Vision Computing 19(9/10):699–707

Hettich S, Bay SD (1998) The UCI KDD archive [http://kdd.ics.uci.edu]. Tech. rep., Department of Information and Computer Science, University of California, Irvine, CA

Krogh A, Vedelsby J (1995) Neural network ensembles, cross validation, and active learning. In: Tesauro G, Touretzky DS, Leen TK (eds) Advances in Neural Information Processing Systems 7, MIT Press, Cambridge, MA, pp 231–238

Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51(2):181–207

Li M, Zhou ZH (2005) SETRED: Self-training with editing. In: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data mining, Hanoi, Vietnam, pp 611–621

Li M, Zhou ZH (2007) Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans 37(6):1088–1098

Liu Y, Yao X (1999a) Ensemble learning via negative correlation. Neural Networks 12(10):1399–1404

Liu Y, Yao X (1999b) Simultaneous training of negatively correlated neural networks in an ensemble. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 29(6):716–725

Lu X, Jain AK (2004) Ethnicity identification from face images. In: Proceedings of SPIE International Symposium on Defense and Security, Kissimmee, FL, pp 114–123

Mallapragada PK, Jin R, Jain AK, Liu Y (2009) Semiboost: Boosting for semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11):2000–2014

Mason L, Bartlett P, Baxter J, Frean M (2000) Functional gradient techniques for combining hypotheses. In: Smola A, Bartlett P, Schölkopf B, Schuurmans D (eds) Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, pp 221–246

McKay R, Abbass H (2001) Analyzing anticorrelation in ensemble learning. In: Proceedings of 2001 Conference on Artificial Neural Networks and Expert Systems, Otago, New Zealand, pp 22–27

Melville P (2005) Creating diverse ensemble classifiers to reduce supervision. PhD thesis, Department of Computer Sciences, University of Texas at Austin, Austin, TX

Melville P, Mooney RJ (2003) Constructing diverse classifier ensembles using artificial training examples. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, pp 505–510

Opitz DW (1999) Feature selection for ensembles. In: Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, FL, pp 379–384

Opitz DW, Shavlik JW (1996) Actively searching for an effective neural network ensemble. Connection Science 8(3&4):337–353

Partridge D, Krzanowski WJ (1997) Software diversity: Practical statistics for its measurement and exploitation. Information and Software Technology 39(10):707–717

Rosen BE (1996) Ensemble learning using decorrelated neural networks. Connection Science 8(3):373–383

Saffari A, Grabner H, Bischof H (2008) SERboost: Semi-supervised boosting with expectation regularization. In: Proceedings of the 10th European Conference on Computer Vision, Marseille, France, pp 588–601

Saffari A, Leistner C, Bischof H (2009) Regularized multi-class semi-supervised boosting. In: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, pp 967–974

Skalak D (1996) The sources of increased accuracy for two proposed boosting algorithms. In: Working Notes of AAAI-96 Workshop on Integrating Multiple Learned Models for Improving & Scaling Machine Learning Algorithms, Portland, OR

Valizadegan H, Jin R, Jain AK (2008) Semi-supervised boosting for multi-class classification. In: Proceedings of the 19th European Conference on Machine Learning, Antwerp, Belgium, pp 522–537

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–83

Wolpert DH (1992) Stacked generalization. Neural Networks 5(2):241–259

Zhang ML, Zhou ZH (2010) Exploiting unlabeled data to enhance ensemble diversity. In: Proceedings of the 10th IEEE International Conference on Data Mining, Sydney, Australia, pp 619–628

Zhang ML, Zhou ZH (in press) CoTrade: Confident co-training with data editing. IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics

Zhou ZH (2009a) Ensemble learning. In: Li SZ (ed) Encyclopedia of Biometrics, Springer, Berlin

Zhou ZH (2009b) When semi-supervised learning meets ensemble learning. In: Proceedings of the 8th International Workshop on Multiple Classifier Systems, Reykjavik, Iceland, pp 529–538

Zhou ZH, Jiang Y (2004) NeC4.5: Neural ensemble based C4.5. IEEE Transactions on Knowledge and Data Engineering 16(6):770–773

Zhou ZH, Li M (2005) Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering 17(11):1529–1541

Zhou ZH, Li M (2010) Semi-supervised learning by disagreement. Knowledge and Information Systems 24(3):415–439

Zhou ZH, Zhan DC, Yang Q (2007) Semi-supervised learning with very few labeled training examples. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence, Vancouver, Canada, pp 675–680

Zhu X (2006) Semi-supervised learning literature survey. Tech. Rep. 1530, Department of Computer Science, University of Wisconsin at Madison, Madison, WI