Front. Electr. Electron. Eng. China, 2010, 5(3)
DOI 10.1007/s11704-009-0000-0

RESEARCH ARTICLE

When Semi-Supervised Learning Meets Ensemble Learning

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China

Abstract Semi-supervised learning and ensemble learning are two important machine learning paradigms. The former attempts to achieve strong generalization by exploiting unlabeled data; the latter attempts to achieve strong generalization by using multiple learners. Although both paradigms have achieved great success during the past decade, they were almost developed separately. In this paper, we advocate that semi-supervised learning and ensemble learning are indeed beneficial to each other, and stronger learning machines can be generated by leveraging unlabeled data and classifier combination.

Keywords Machine learning; semi-supervised learning; ensemble learning

1 Introduction

Constructing learning systems with strong generalization ability is one of the ultimate goals of machine learning. During the past decades two learning paradigms, *semi-supervised learning* and *ensemble learning*, have achieved great success. The former [11, 57, 60] tries to achieve strong generalization by exploiting unlabeled data, while the latter [50] tries to achieve strong generalization by using the combination of multiple learners. It is noteworthy that, however, these two paradigms were developed almost in parallel. Only a few studies touched both of them [6, 13, 27, 28, 40, 55] while most were semi-supervised boosting methods [6, 13, 28, 40], and very few for the fuse of other kinds of semi-supervised and ensemble learning [27, 55].

This phenomenon has been attributed by [51] to the fact that the semi-supervised learning community and the ensemble learning community have different philosophies. From the view of the semi-supervised learning community, it seems that using unlabeled data to boost the learning performance is good enough, and so there is no need to involve multiple learners; while from the view of the ensemble learning community, it seems that using multiple learners can do all the things and therefore there is no need to consider unlabeled data.

It has been advocated in [51] that semi-supervised learning and ensemble learning are indeed beneficial to each other, and stronger learning machines can be generated by leveraging unlabeled data and classifier combination. In this article, using *disagreement-based semi-supervised learning* [57] as an example, we will provide more details to the arguments in [51] on why it is good to leverage unlabeled data and classifier combination.

We will start by a brief background introduction in Section 2. Then, we will discuss on why classifier combination can be helpful to semi-supervised learning in Section 3, and why unlabeled data can be helpful to ensemble learning in Section 4. Finally we conclude the article in Section 5.

2 Background

2.1 Ensemble Learning

Ensemble methods train multiple learners to solve the same problem. In contrast to ordinary learning approaches which try to construct one learner from training data, ensemble methods try to construct a set of learners and combine them to use. Ensemble learning is also called as committee-based learning, or learning with multiple classifier systems.

The generalization ability of an ensemble is usually much stronger than that of its component learners. Ensemble learning is appealing mostly because that it is able to boost weak learners which are slightly better than random guess to strong learners which can make very accurate predictions. An ensemble is typically constructed

Received month dd.yyyy; accepted month dd.yyyy

E-mail: *zhouzh@nju.edu.cn*

in two steps. First, a number of component learners are generated; then, the component learners are combined for prediction. Generally, to get a good ensemble, the component learners should be as more accurate as possible, and as more diverse as possible [24]. However, how to measure and control the diversity remains an open problem though recently there are some promising advances [10, 58].

Famous ensemble learning methods including Bagging [8], Boosting [15], Random Subspace [19], Random Forest [9], etc.

2.2 Semi-Supervised Learning

In many real applications it is difficult to get a large amount of labeled training examples although there may exist abundant unlabeled data, since labeling the unlabeled instances requires human effort and expertise. Exploiting unlabeled data to help improve the learning performance has become a very hot topic during the past decade. There are three major techniques for this purpose [49], i.e., semi-supervised learning, transductive learning and active learning.

Semi-supervised learning [11, 57, 60] deals with methods for exploiting unlabeled data in addition to labeled data automatically to improve learning performance, where no human intervention is assumed. Transductive learning [41] also tries to exploit unlabeled data automatically, but it assumes that the unlabeled examples are exactly the test examples. Active learning [37] deals with methods which assume that the learner has some control over the input space, and the goal is to minimize the number of queries from human experts on ground-truth labels for building a strong learner. In this article we will focus on semi-supervised learning, and regards transductive learning as a special kind of semi-supervised learning since both try to exploit unlabeled data without human intervention.

Many semi-supervised learning algorithms have been developed. Roughly speaking, they can be categorized into four categories, i.e., generative methods [30, 34, 38, 44], S3VMs (Semi-Supervised Support Vector Machines) [12, 17, 22, 25], graph-based methods [3–5, 48, 61], and disagreement-based methods [57].

2.3 Disagreement-based Semi-Supervised Learning

Disagreement-based semi-supervised learning [57] generates multiple learners, lets them collaborate to exploit unlabeled instances, and maintains a large disagreement between the learners to enable the learning process to continue. We pay more attention to this kind of semi-supervised learning because its learning process involves multiple learners and thus provides a good example for studying whether it is beneficial to leverage unlabeled data and classifier combination.

Research on disagreement-based semi-supervised learning started from Blum and Mitchell's seminal work on co-training [7]. They considered the situation where data have two *sufficient and redundant* views (i.e., two attribute sets each of which contains sufficient information for constructing a strong learner and is conditionally independent to the other attribute set given the class label). The algorithm trains a learner from each view using the original labeled data. Each learner selects and labels some high-confident unlabeled instances for its peer. Then, each learner is refined using the newly labeled examples provided by its peer. The whole process repeats until no learner changes or a pre-set number of learning rounds is reached.

Blum and Mitchell [7] analyzed the effectiveness of co-training and disclosed that if the two views are conditionally independent, the predictive accuracy of an initial weak learner can be boosted to arbitrarily high using unlabeled data by co-training. Dasgupta et al. [14] showed that when the two views are sufficient and conditionally independent, the generalization error of co-training is upper-bounded by the disagreement between the two classifiers. Later, Balcan et al. [2] indicated that if a PAC learner can be obtained on each view, the conditional independence assumption or even the weak independent assumption [1] is unnecessary, and a weaker assumption of "expansion" of the underlying data distribution is sufficient for iterative co-training to succeed.

The requirement of two sufficient and redundant views is too luxury to be satisfied in most real-world tasks, since there is generally only one attribute set. Thus, the applicability of the standard co-training is limited though Nigam and Ghani [33] showed that if there exist large redundancy in the attribute set, co-training can be enabled through view splitting. To deal with single view data, Goldman and Zhou [16] proposed a method which trains two learners by using different learning algorithms. The method requires each classifier be able to partition the instance space into equivalence classes, and uses cross validation to estimate the confidences of the two learners as well as the equivalence classes.

Zhou and Li [55] proposed the *tri-training* method, which requires neither two views nor special learning algorithms. This method uses three learners and avoids estimating the predictive confidence explicitly. It employs "majority teach minority" strategy for the semi-supervised learning process, that is, if two learners agree on an unlabeled instance yet the third learner disagrees, the two learners will label this instance for the third learner. Moreover, classifier combination is exploited to improve generalization. Later, Li and Zhou [27] proposed the *co-forest* method by extending tri-training to include more learners. In co-forest, each learner is improved with unlabeled instances labeled by the ensemble consists of all the other learners, and the final prediction is made by the ensemble of all learners. Zhou and

Li [54, 56] proposed the first semi-supervised regression algorithm Coreg which employs two kNN regressors facilitated with different distance metrics. This algorithm does not require two views either. Later it was extended to a semi-supervised ensemble method for time series prediction with missing data [31].

Previous theoretical studies [2, 7, 14] focused on situations with two views, and could not explain why single-view methods can work. Wang and Zhou [42] presented a theoretical analysis which discloses that the key for disagreement-based approaches to succeed is that there exists a large diversity between the learners, while it is unimportant whether the diversity is achieved by using two views, or by using two learning algorithms, or from other channels.

Disagreement-based semi-supervised learning approaches have been applied to many real-world tasks, such as natural language processing [21, 35, 36, 39], document retrieval [26], spam detection [29], email answering [23], mammogram microcalcification detection [27], etc. An effective method which combines disagreement-based semi-supervised learning with active learning for image retrieval has been developed in [52, 53], with a theoretical analysis presented recently in [43].

3 The Helpfulness of Classifier Combination to Semi-Supervised Learning

Here we briefly introduce some of our theoretical results on the helpfulness of classifier combination to semi-supervised learning.

Given data set $\mathcal{S} = \mathcal{L} \cup \mathcal{U}$, where $\mathcal{L} = \{(x_1, y_1), \cdots, (x_l, y_l)\} \subset \mathcal{X} \times \mathcal{Y}$ are labeled data and $\mathcal{U} = \{x_{l+1}, x_{l+2}, \cdots, x_n\} \subset \mathcal{X}$ are unlabeled data. $\mathcal{Y} = \{-1, +1\}$; $\mathcal{X}$ is with distribution $\mathcal{D}$. Let $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ denote the hypothesis space. Assume that $|\mathcal{H}|$ is finite, and $\mathcal{D}$ is generated by the ground truth $h^* \in \mathcal{H}$. For a finite sample, it is hard to achieve h^* over $\mathcal{S}$. Suppose we obtain a learner $h^i \in \mathcal{H}$ from $\mathcal{S}$, which is somewhat different from h^*. Let $d(h^i, h^*)$ denote the disagreement between h^i and h^*, then

$$d(h^i, h^*) = Pr_{x \in \mathcal{D}}[h^i(x) \neq h^*(x)].$$

Let ϵ bound the generalization error of the learners what we wish to achieve finally. That is, if $d(h^i, h^*) = Pr_{x \in \mathcal{D}}[h^i(x) \neq h^*(x)] < \epsilon$, we say that we have got a desired learner; otherwise we say that the learner h^i is bad. We wish to have a high probability to achieve a good learner. The learning process is said to do probably approximately correct learning of h^* if and only if $Pr[d(h^i, h^*) \geq \epsilon] \leq \delta$, the disagreement between the ground truth h^* and the hypothesis h^i should be small (less than ϵ) with high probability (larger than $1 - \delta$).

We consider the following disagreement-based semi-supervised learning process:

Process: *At first, we train two initial learners h_1^0 and h_2^0 using $\mathcal{L}$ which contains l labeled examples. Then, h_1^0 selects u number of unlabeled instances from $\mathcal{U}$ to label, and puts these newly labeled examples into the data set σ_2 which contains all the examples in $\mathcal{L}$; at the same time, h_2^0 selects u number of unlabeled instances from $\mathcal{U}$ to label, and puts these newly labeled examples into the data set σ_1 which contains all the examples in $\mathcal{L}$. Then, h_1^1 and h_2^1 are trained from σ_1 and σ_2, respectively. After that, h_1^1 selects u number of unlabeled instances to label, and uses these newly labeled examples to update σ_2; while h_2^1 also selects u number of unlabeled instances to label, and uses these newly labeled examples to update σ_1. Such a process is repeated for a pre-set number of learning rounds.*

If the above process is able to boost the performance to arbitrarily high by using unlabeled data, as believed previously, then there is of course no need to exploit classifier combination. However, contrasting to previous believes, a typical empirical observation looks like Fig. 1, where the errors of the two classifiers could not drop further after a number of learning rounds, far from the expected ideal case where the error were expected to drop to arbitrarily low.

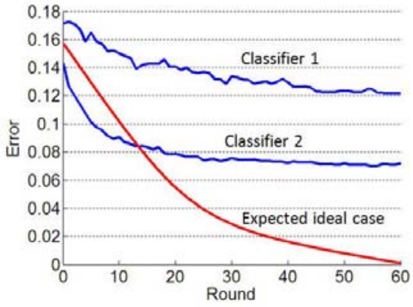

Fig. 1 An illustration of typical empirical observation.

Wang and Zhou [42] proved that the occur of such a phenomenon is not occasional, because the key condition for disagreement-based learning process to continue is that the two learners have a large diversity; while during the learning process, the two learners will become more and more similar since they keep on teaching each other, and the consequence is that after some rounds the diversity between them could not allow the learning process to continue effectively.

Thus, we know that it is hard to improve the performance to arbitrarily high by using unlabeled data, and so we are interested in whether the use of classifier combination can be helpful for a further improvement.

In the following we just briefly introduce some of our theoretical results. Proofs and more details can be found in a longer version of [42].

3.4 Combination Can be Improved Even When Individual classifiers Could not be Improved Further

First, we can prove that even when the individual learners could not improve the performance any more, classifier combination is still possible to improve generalization further. We start by Theorem 1 which bounds the error of the individual learners.

Theorem 1 *Given the initial labeled data set $\mathcal{L}$ which is clean, and assuming that the size of $\mathcal{L}$ is sufficient to learn two learners h_1^0 and h_2^0 whose upper bound of the generalization error is $a_0 < 0.5$ and $b_0 < 0.5$, respectively, with high probability (more than $1-\delta$) in the PAC model, i.e., $l \geq max[\frac{1}{a_0}ln\frac{|\mathcal{H}|}{\delta}, \frac{1}{b_0}ln\frac{|\mathcal{H}|}{\delta}]$. Then h_1^0 selects u number of unlabeled instances from $\mathcal{U}$ to label and puts them into σ_2 which contains all the examples in $\mathcal{L}$, and then h_2^1 is trained from σ_2 by minimizing the empirical risk. If $lb_0 \leq e\sqrt[M]{M!} - M$, then*

$$Pr[d(h_2^1, h_*) \geq b_1] \leq \delta.$$

Here $M = ua_0$ and $b_1 = \max[\frac{lb_0+ua_0-ud(h_1^0,h_2^1)}{l}, 0]$.

Here for simplicity, we consider the combination strategy of *weighted voting* in [32]. In detail, for some instance x, if the confidence of h_j^i $(j = 1, 2)$ is larger than that of h_{3-j}^i, the label of the instance x will be assigned according to $h_j^i(x)$. Let S^i denote the set of the instances in which $h_j^i(x) \neq h_{3-j}^i(x)$ and let h_{com}^i denote the combination of the classifiers h_j^i and h_{3-j}^i. For the convenience of analysis, we assume that $Pr_{x \in S^i}[h_{com}^i(x) \neq h^*(x)] = \gamma_i$. γ_i indicates the property of h_{com}^i in the disagreement region of the individual classifiers. Thus the error rate of the combination h_{com}^i can be expressed as

$$\begin{aligned} & Pr\left[h_{com}^i(x) \neq h^*(x)\right] \\ = & \frac{1}{2}\left(Pr\left[h_1^i(x) \neq h^*(x)\right] + Pr\left[h_2^i(x) \neq h^*(x)\right]\right) \\ & +(\gamma_i - \frac{1}{2})Pr\left[h_1^i(x) \neq h_2^i(x)\right] \ . \end{aligned} \tag{1}$$

Suppose that the decrease of the diversity between the two classifiers can be expressed as a non-negative function τ with parameters of the classifier h_j^i $(j = 1, 2)$ and the set of u newly labeled instances $S_{h_j^i}^u$ provided by h_j^i. This implies that

$$\begin{aligned} & d(h_1^{i-1}, h_2^{i-1}) - d(h_1^{i-1}, h_2^i) = \tau(h_1^{i-1}, S_{h_1^{i-1}}^u) \\ & d(h_1^{i-1}, h_2^{i-1}) - d(h_1^i, h_2^{i-1}) = \tau(h_2^{i-1}, S_{h_2^{i-1}}^u) \\ & d(h_1^{i-1}, h_2^{i-1}) - d(h_1^i, h_2^i) \\ = & \tau(h_1^{i-1}, S_{h_1^{i-1}}^u) + \tau(h_2^{i-1}, S_{h_2^{i-1}}^u). \end{aligned}$$

It is not difficult to see that $d(h_1^{i-1}, h_2^i) + d(h_1^i, h_2^{i-1}) = d(h_1^{i-1}, h_2^{i-1}) + d(h_1^i, h_2^i)$. For the convenience of discussion, suppose $\gamma_i = \gamma$ in the i-th round. We have Theorem 2 on the performance of the classifier combination.

Theorem 2 *If $d(h_1^1, h_2^1) > \frac{ua_0+ub_0+\left(l(1-2\gamma)-u\right)d(h_1^0,h_2^0)}{u+l(1-2\gamma)}$ and $u > l$, then*

$$Pr[h_{com}^1(x) \neq h^*(x)] < Pr[h_{com}^0(x) \neq h^*(x)].$$

Based on Theorems 1 and 2, we have the following theorem which indicates that even when the performances of both individual classifiers are no longer improved, the performance of the combination of the individual classifiers could still be improved further.

Theorem 3 *If $d(h_1^0, h_2^0) > a_0 > b_0$ and $\gamma \geq \frac{1}{2} + \frac{u\left(a_0+b_0-d(h_1^0,h_2^0)\right)}{2ld(h_1^0,h_2^0)}$, even when $Pr[h_j^1(x) \neq h^*(x)] \geq Pr[h_j^0(x) \neq h^*(x)]$ $(j = 1, 2)$, $Pr[h_{com}^1(x) \neq h^*(x)]$ is still less than $Pr[h_{com}^0(x) \neq h^*(x)]$.*

3.5 Combination Can be Better than Individual Classifiers in Prediction

Without loss of generality, assume that $a_0 > b_0$. For any instance x, let $\phi_j^i(x) : X \to [0, 1]$ $(j = 1, 2)$ denote the confidence of the prediction $h_j^i(x)$, so the combination h_{com}^i of the individual learners according to *weighted voting* in [32] can be formulated as

$$h_{com}^i(x) = \begin{cases} h_1^i(x) \text{ if } \phi_1^i(x) > \phi_2^i(x) \\ h_2^i(x) \text{ otherwise }. \end{cases} \tag{2}$$

We define the confidence risk $CR(\cdot)$ and the confidence gain $CG(\cdot)$ of h_j^i as:

$$CR(h_j^i) = \frac{\int_{h_j^i(x) \neq h^*(x)} \phi_j^i(x)p(x)dx}{\int_{h_j^i(x) \neq h^*(x)} p(x)dx}$$

$$CG(h_j^i) = \frac{\int_{h_j^i(x) = h^*(x)} \phi_j^i(x)p(x)dx}{\int_{h_j^i(x) = h^*(x)} p(x)dx}$$

It is not difficult to see that if $CR(h_1^i) < CG(h_2^i)$ and $CR(h_2^i) < CG(h_1^i)$ hold, h_{com}^i in Eq. 2 will correctly classify the instances in the disagreement region with large probability.

Suppose that the distribution $\mathcal{D}$ over the example space $\mathcal{X}$ is the uniform distribution. Assume that $\phi_1^0(x)$ is uniformly distributed over $[CR(h_1^0) - \alpha_1, CR(h_1^0) + \alpha_1]$ for the instance set in which $h_1^0(x) \neq h^*(x)$, and $\phi_1^0(x)$ is uniformly distributed over $[CG(h_1^0) - \beta_1, CG(h_1^0) + \beta_1]$ for the instance set in which $h_1^0(x) = h^*(x)$; similarly, $\phi_2^0(x)$ is uniformly distributed over $[CR(h_2^0) - \alpha_2, CR(h_2^0) + \alpha_2]$ for the instance set in which $h_2^0(x) \neq h^*(x)$, and $\phi_2^0(x)$ is uniformly distributed over $[CG(h_2^0) - \beta_2, CG(h_2^0) + \beta_2]$ for the instance set in which $h_2^0(x) = h^*(x)$. Without loss of generality, assume $CR(h_1^0) + \alpha_1 > CG(h_2^0) - \beta_2$ and $CR(h_2^0) + \alpha_2 > CG(h_1^0) - \beta_1$.

We have the following theorem which indicates that the combination can be better than individual classifiers in prediction.

Theorem 4 *Suppose the combination h^0_{com} of the individual learners h^0_1 and h^0_2 is generated according to Eq. 2. If*

$$a_0\alpha_2\big(CR(h^0_1) - CG(h^0_2)\big) + b_0\alpha_1\big(CR(h^0_2) - CG(h^0_1)\big)$$
$$< \alpha_1\alpha_2\big(d(h^0_1, h^0_2) - 2a_0\big) - a_0\alpha_2\beta_2 - b_0\alpha_1\beta_1,$$

then the error rate of h^0_{com} is less than $\min[a_0, b_0]$.

4 The Helpfulness of Unlabeled Data to Ensemble Learning

4.6 Unlabeled Data Can Enable Ensemble Learning When There are Very Few Labeled Data

Generally, a lot of labeled training examples are needed for constructing a strong ensemble. In some real applications, however, the number of labeled training examples may be too few to launch a successful ensemble learning. Under such situations, unlabeled data may be helpful for enabling ensemble learning.

Zhou et al. [59] showed that, when the assumption of the classical disagreement-based semi-supervised learning method, co-training, is hold, it is possible to exploit unlabeled data to help enrich the labeled training examples. This is feasible even when there is only one labeled example. After such a process, standard ensemble methods can be applied. In the following we briefly introduce the method presented in [59].

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the two sufficient and redundant views, that is, two independent attribute sets each contain sufficient information for constructing a strong classifier. Let $(\langle \boldsymbol{x}, \boldsymbol{y}\rangle, c)$ denote a labeled example where $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ are the two portions of the example, and c is the label. For simplifying the discussion, assume that $c \in \{0, 1\}$ where 0 and 1 denote negative and positive classes, respectively. Given $(\langle \boldsymbol{x}_0, \boldsymbol{y}_0\rangle, 1)$ and a large number of unlabeled instances $\mathcal{U} = \{(\langle \boldsymbol{x}_i, \boldsymbol{y}_i\rangle, c_i)\}$ $(i = 1, \cdots, l-1$; c_i is unknown), we hope to exploit $\mathcal{U}$ to enrich the labeled examples.

Since the two views are sufficient and independent, some projections on these two views should have strong correlation with the ground truth. If the correlated projections of these two views can be identified, they can help induce the labels of some unlabeled instances.

The correlation between the projections on the two views can be identified by the canonical correlation analysis [20] or its kernel extension [18], and a number of correlated pairs of projections can be identified. If the two views are really conditionally independent given the class label, the most strongly correlated pair of projections should be in accordance with the ground-truth. In real applications, however, the conditional independence rarely holds and therefore, information conveyed by the other pairs of correlated projections should not be omitted. The Oltv method [59] takes a simple strategy as follows to use these information.

Let m denote the number of pairs of correlated projections that have been identified, an instance $\langle \boldsymbol{x}^*, \boldsymbol{y}^*\rangle$ can be projected into $\langle P_j(\boldsymbol{x}^*), P_j(\boldsymbol{y}^*)\rangle$ $(j = 1, 2, \cdots, m)$. Then, in the jth projection, the similarity between an original unlabeled instance $\langle \boldsymbol{x}_i, \boldsymbol{y}_i\rangle$ $(i = 1, 2, \cdots, l-1)$ and the original labeled instance $\langle \boldsymbol{x}_0, \boldsymbol{y}_0\rangle$, $sim_{i,j}$, can be measured. Considering that $\langle \boldsymbol{x}_0, \boldsymbol{y}_0\rangle$ is a positive instance, $\rho_i = \sum_{j=1}^{m} \lambda_j sim_{i,j}$ delivers the confidence of $\langle \boldsymbol{x}_i, \boldsymbol{y}_i\rangle$ being a positive instance, where λ_j is an coefficient reflecting the strength of the correlation solved by the method. Thus, several unlabeled examples with the highest and lowest ρ values can be selected, respectively, as the extra positive and negative examples. The number of labeled training examples is thus increased.

The quality of the additional labeled training examples derived by the Oltv method is much better than that derived by using strategies such as k nearest neighbor in the original feature space (e.g., using the k unlabeled instances nearest to $\langle \boldsymbol{x}_0, \boldsymbol{y}_0\rangle$ as additional positive examples while the k farthest unlabeled instances as additional negative ones). Let *reliability* be b/a if the labels for a unlabeled instances have been induced among which b of them are correct. As shown in Fig. 2 [59], where δ is an coefficient related to the number of additional labeled examples induced, the reliability of the kNN strategy is far worse than that of Oltv. On all experimental data sets, the reliability of Oltv is always higher than 80% and even often higher than 90%.

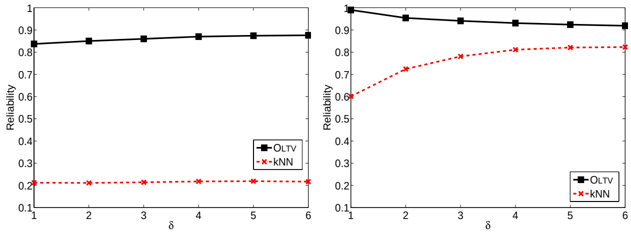

Fig. 2 The reliability of Oltv and kNN on two data sets

Thus, the labeled training data can be significantly enriched by exploiting unlabeled data, and thus ensemble learning can be enabled.

4.7 Unlabeled Data Can Help Enhance Diversity for Ensemble Learning

It is well-known that the more accurate and the more diverse the component classifiers, the better the ensemble [24]. Standard ensemble methods work under supervised setting, trying to achieve a high average accuracy and a high diversity for component classifiers by using the labeled training data. It is noteworthy, however, pursuing a high accuracy and a high diversity may suffer from a dilemma. For example, for two classifiers that 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, the increase of diversity needs to sacrifice the accuracy of one classifier.

When unlabeled data are considered, however, it may be found that these two classifiers indeed 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 for taking either the ensemble consists of (A, B) or the ensemble consists 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 consists of (C, D) would have good chance to be better. So, in contrast to standard ensemble methods that 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 the following we briefly introduce the work presented in [46, 47].

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, let $\tilde{\mathcal{L}} = \{\boldsymbol{x}_i | 1 \leq i \leq L\}$ denote the unlabeled data set derived from $\mathcal{L}$ by neglecting the label information. Assume that the classifier ensemble is composed of m component classifiers $\{f_k | 1 \leq k \leq m\}$, each taking the form $f_k : \mathcal{X} \to [-1, +1]$. 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 idea of the Udeed method 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 generates the classifier ensemble $\boldsymbol{f} = (f_1, f_2, \cdots, f_m)$ by minimizing the loss function

$$V(\boldsymbol{f}, \mathcal{L}, \mathcal{D}) = V_{emp}(\boldsymbol{f}, \mathcal{L}) + \gamma \cdot V_{div}(\boldsymbol{f}, \mathcal{D}), \qquad (3)$$

where $V_{emp}(\boldsymbol{f}, \mathcal{L})$ corresponds to the *empirical loss* of $\boldsymbol{f}$ on the labeled data set $\mathcal{L}$, $V_{div}(\boldsymbol{f}, \mathcal{D})$ corresponds to the *diversity loss* of $\boldsymbol{f}$ on a specified data set $\mathcal{D}$ (e.g., $\mathcal{D} = \mathcal{U}$), and γ is the parameter trades off the importance of the two terms.

The Udeed method calculates $V_{emp}(\boldsymbol{f}, \mathcal{L})$ in Eq. 3 according to

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

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

It calculates $V_{div}(\boldsymbol{f}, \mathcal{D})$ according to

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

where

$$d(f_p, f_q, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum\nolimits_{\boldsymbol{x} \in \mathcal{D}} f_p(\boldsymbol{x}) f_q(\boldsymbol{x}).$$

$|\mathcal{D}|$ returns the cardinality of data set $\mathcal{D}$. Intuitively, $d(f_p, f_q, \mathcal{D})$ represents the *prediction difference* between any pair of component classifiers on a specified data set $\mathcal{D}$. Note that the prediction difference is calculated based on the real output $f(\boldsymbol{x})$ instead of the signed output sign$[f(\boldsymbol{x})]$. In this way, the *prediction confidence* of each classifier is considered.

Then, Udeed aims to find the target model $\boldsymbol{f}^*$ which minimizes the loss function in Eq.(3):

$$\boldsymbol{f}^* = \arg\min\nolimits_{\boldsymbol{f}} V(\boldsymbol{f}, \mathcal{L}, \mathcal{D})$$

The Udeed method is implemented by using logistic regression to realize the component classifiers. Empirical study on a broad range of data sets show that the Udeed method is even superior to famous ensemble methods such as Bagging and AdaBoost for prediction.

Moreover, the diversity of the Udeed ensemble was studied using four different diversity measures, i.e., the *Disagreement measure* (DIS), the *Double-fault measure* (DF), the *Entropy measure* (ENT) and the *Coincident failure diversity* (CFD). The diversity of the initial ensemble (before using unlabeled data) and that of the final ensemble (after using unlabeled data) were compared by pairwise t-tests on each data set. A win/loss was recorded if the final ensemble achieved a significantly higher/lower diversity than the initial one, and otherwise a tie was recorded. The win/tie/loss counts in terms of the four diversity measures under various ensemble sizes are summarized in Table 1, which clearly shows that exploiting unlabeled data is helpful for enhancing the diversity for ensemble learning.

5 Discussions and Concluding Remarks

5.8 Summary

Semi-supervised learning and ensemble learning are two well-developed paradigms for constructing strong learning machines. Possibly due to the different philosophies of the semi-supervised learning community and the ensemble learning community, though both have achieved great success during the past decade, they were almost developed separately.

Table 1 Win/tie/loss counts for FINAL ensemble against INITIAL ensemble in terms of four diversity measures.

	FINAL ensemble vs. INITIAL ensemble											
	ensemble size = 20				ensemble size = 50				ensemble size = 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**

In this article, we provide more details to the argument in [51] that semi-supervised learning and ensemble learning are indeed beneficial to each other. We show that:

- Classifier combination is helpful to semi-supervised learning. There are at least two reasons: 1) the performance of classifier combination can be improved further even when the individual classifiers could not be improved using unlabeled data; 2) the classifier combination can reach a good performance earlier than individual classifiers.
- Unlabeled data are helpful to ensemble learning. There are at least two reasons: 1) when there are very few labeled training examples, unlabeled data can help to enable ensemble learning by inducing additional labeled training examples; 2) unlabeled data can be exploited to help enhance the diversity of component classifiers.

Indeed we can provide even more reasons for considering classifier combination in semi-supervised learning (e.g., classifier combination can be employed to improve the stability of exploiting unlabeled data), and more reasons for considering unlabeled data in ensemble learning (e.g., unlabeled data can be used in dimensionality reduction to help ensemble learning handle high-dimensional data).

Note that some of our arguments were made based on taking disagreement-based semi-supervised learning approaches as examples. However, most of them are possible to be generalized to other kinds of semi-supervised learning and ensemble learning approaches.

5.9 The Special Role of Disagreement-based Methods

If we look back the developmental trails of ensemble methods, we may notice that there were three threads of early contributions which led to the current ensemble learning area, that is, “combining classifiers”, “ensembles of weak learners” and “mixture-of-experts”.

“Combining classifiers” was mostly studied in the pattern recognition community. In this thread of work, researchers generally work on strong classifiers, and try to design powerful *combining rules* to get stronger combined classifier. As the consequence, this thread of work has accumulated deep understanding on the design and use of different combining rules. “Ensembles of weak learners” was mostly studied in the machine learning

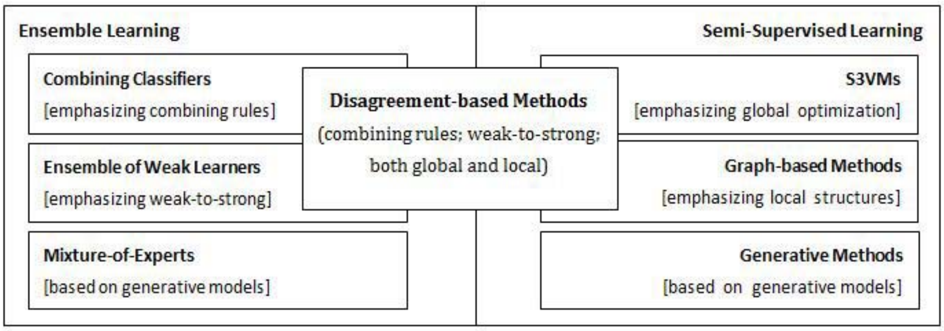
Ensemble Learning
Semi-Supervised Learning
Combining Classifiers
[emphasizing combining rules]
Ensemble of Weak Learners
[emphasizing weak-to-strong]
Mixture-of-Experts
[based on generative models]
Disagreement-based Methods
(combining rules; weak-to-strong; both global and local)
S3VMs
[emphasizing global optimization]
Graph-based Methods
[emphasizing local structures]
Generative Methods
[based on generative models]

Fig. 3 The role of disagreement-based methods

community. In this thread of work, researchers generally work on weak classifiers that are just slightly better than random guess, and try to design powerful learning algorithms to boost the performance from weak to strong. This thread of work has led to the born of famous ensemble methods such as AdaBoost, Bagging, etc., and theoretical understanding on why and how weak learners can be boosted to strong ones. “Mixture-of-experts” was mostly studied in the neural networks community. In this thread of work, researchers generally consider a “divide-and-conquer” strategy, trying to learn a mixture of parametric models jointly and use combining rules to get an overall solution. There is a nice article discussing on the relation between “combining classifiers” and “mixture-of-experts” [45].

As mentioned in Section 2, currently there are four major categories of semi-supervised learning methods. One category, i.e., generative methods, are very similar to “mixture-of-models”, both assuming a generative model and using the EM process for parameter estimation; the technical essential is almost the same, and so in the following we do not talk more about this category.

If we want to enhance the connection between ensemble learning and semi-supervised learning, and if we consider the major categories of semi-supervised learning methods together with the major threads of ensemble studies, it is interesting to see that the disagreement-based methods play a special role in the interaction between ensemble learning and semi-supervised learning, as shown in Fig. 3.

It is evident that the category of disagreement-based methods exhibits a natural intersection between ensemble learning and semi-supervised learning, since it considers multiple learners as well as the exploitation of unlabeled data. More importantly, different threads of ensemble learning and different categories of semi-supervised learning provide generous support to the study of disagreement-based methods. On one aspect, important ingredients of ensemble studies, such as the design and understanding of different combining rules, and the theoretical and algorithmic studies on how to drive weak learners to strong learners, are all beneficial to the development of disagreement-based methods, and indeed some have already been utilized; these properties were rarely touched by other categories of semi-supervised learning methods. On the other aspect, important ingredients of semi-supervised studies, such as the exploitation of global optimization techniques and the consideration of local structures of data distribution, are all beneficial to the development of disagreement-based methods, and some have already been utilized; these properties were rarely emphasized by ensemble studies.

Overall, disagreement-based methods provide a good vessel to accommodate research advantages from ensemble studies and semi-supervised studies. By recognizing its special role, we believe that in the future there will be more interesting work on disagreement-based methods, for the fuse of advantages from semi-supervised learning and ensemble learning to generate much stronger learning machines.

6 Acknowledgments

This is a summary of the invited talk at IScIDE’2010. The author wants to thank Prof. Lei Xu for invaluable comments and suggestions, especially on the special role of disagreement-based methods. This research was supported by the National Science Foundation of China (60635030), the Jiangsu Science Foundation (BK2008018) and the National Fundamental Research Program of China (2010CB327903).

References

1. S. Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, PA, 2002.
2. M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances*

in Neural Information Processing Systems 17, pages 89–96. MIT Press, Cambridge, MA, 2005.

3. M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1-3):209–239, 2004.
4. M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 17–24, Savannah, Barbados, 2005.
5. M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.
6. 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*, pages 289–296, Edmonton, Canada, 2002.
7. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.
8. L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
9. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
10. G. Brown. An information theoretic perspective on multiple classifier systems. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 344–353, Reykjavik, Iceland, 2009.
11. O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
12. O. Chapelle and A. Zien. Semi-supervised learning by low density separation. In *proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 57–64. Savannah Hotel, Barbados, 2005.
13. F. d'Alché-Buc, Y. Grandvalet, and C. Ambroise. Semi-supervised MarginBoost. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 553–560. MIT Press, Cambridge, MA, 2002.
14. S. Dasgupta, M. Littman, and D. McAllester. PAC generalization bounds for co-training. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 375–382. MIT Press, Cambridge, MA, 2002.
15. Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
16. S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proceedings of the 17th International Conference on Machine Learning*, pages 327–334, San Francisco, CA, 2000.
17. Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 529–536. MIT Press, Cambridge, MA, 2005.
18. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
19. T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.
20. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(4):321–377, 1936.
21. R. Hwa, M. Osborne, A. Sarkar, and M. Steedman. Corrected co-training for statistical parsers. In *Working Notes of the ICML'03 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington, DC, 2003.
22. T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, Bled, Slovenia, 1999.
23. M. Kockelkorn, A. Lüneburg, and T. Scheffer. Using transduction and multi-view learning to answer emails. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 266–277, Cavtat-Dubrovnik, Croatia, 2003.
24. A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 231–238. MIT Press, Cambridge, MA, 1995.
25. N. D. Lawrence and M. I. Jordan. Semi-supervised learning via Gaussian processes. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 753–760. MIT Press, Cambridge, MA, 2005.
26. M. Li, H. Li, and Z.-H. Zhou. Semi-supervised document retrieval. *Information Processing & Management*, 45(3):341–355, 2009.
27. M. Li and Z.-H. Zhou. 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, 2007.
28. 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*, 31(11):2000–2014, 2009.
29. D. Mavroeidis, K. Chaidos, S. Pirillos, D. Christopoulos, and M. Vazirgiannis. Using tri-training and support vector machines for addressing the ECML-PKDD 2006 Discovery Challenge. In *Proceedings of ECML-PKDD 2006 Discovery Challenge Workshop*, pages 39–47, Berlin, Germany, 2006.
30. D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In M. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 571–577. MIT Press, Cambridge, MA, 1997.
31. T. A. Mohamed, N. El Gayar, and A. F. Atiya. A co-training approach for time series prediction with missing data. In *Proceedings of the 7th International Workshop on Multiple Classifier Systems*, pages 93–102, Prague, Czech, 2007.
32. I. Muslea, S. Minton, and C. A. Knoblock. Active learn-

ing with multiple views. *Journal of Artificial Intelligence Research*, 27:203–233, 2006.

33. K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the 9th ACM International Conference on Information and Knowledge Management*, pages 86–93, Washington, DC, 2000.
34. K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.
35. D. Pierce and C. Cardie. Limitations of co-training for natural language learning from large data sets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 1–9, Pittsburgh, PA, 2001.
36. A. Sarkar. Applying co-training methods to statistical parsing. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 95–102, Pittsburgh, PA, 2001.
37. B. Settles. Active learning literature survey. Technical Report 1648, Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI, 2009. http://pages.cs.wisc.edu/∼bsettles/pub/settles.activelearning.pdf.
38. B. Shahshahani and D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994.
39. M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. Bootstrapping statistical parsers from small data sets. In *Proceedings of the 11th Conference on the European Chapter of the Association for Computational Linguistics*, pages 331–338, Budapest, Hungary, 2003.
40. 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*, pages 522–537, Antwerp, Belgium, 2008.
41. V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.
42. W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *Proceedings of the 18th European Conference on Machine Learning*, pages 454–465, Warsaw, Poland, 2007.
43. W. Wang and Z.-H. Zhou. On multi-view active learning and the combination with semi-supervised learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1152–1159, Helsinki, Finland, 2008.
44. L. Xu. Bayesian ying yang system and theory as a unified statistical learning approach: (i) unsupervised and semi-unsupervised learning. In S. Amari and N. Kassabov, editors, *Brain-like Computing and Intelligent Information Systems*, pages 241–274. Springer, Berlin, 1997.
45. L. Xu and S. Amari. Combining classifiers and learning mixture-of-experts. In J. R. R. Dopico, J. Dorado, and A. Pazos, editors, *Encyclopedia of Artificial Intelligence*, pages 318–326. IGI, Berlin, 2009.
46. M.-L. Zhang and Z.-H. Zhou. Classifier ensemble with unlabeled data. CORR abs/0909.3593, 2009.
47. M.-L. Zhang and Z.-H. Zhou. Exploiting unlabeled data to enhance ensemble diversity. In *Proceedings of the 9th IEEE International Conference on Data Mining*, Sydney, Australia, 2010.
48. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, 2004.
49. Z.-H. Zhou. Learning with unlabeled data and its application to image retrieval. In *Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence*, pages 5–10, Guilin, China, 2006 (keynote speech).
50. Z.-H. Zhou. Ensemble learning. In S. Z. Li, editor, *Encyclopedia of Biometrics*, pages 270–273. Springer, Berlin, 2009.
51. Z.-H. Zhou. When semi-supervised learning meets ensemble learning. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 529–538, Reykjavik, Iceland, 2010 (invited plenary talk).
52. Z.-H. Zhou, K.-J. Chen, and H.-B. Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 24(2):219–244, 2006.
53. Z.-H. Zhou, K.-J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. In *Proceedings of the 15th European Conference on Machine Learning*, pages 525–536, Pisa, Italy, 2004.
54. Z.-H. Zhou and M. Li. Semi-supervised regression with co-training. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 908–913, Edinburgh, Scotland, 2005.
55. Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005.
56. Z.-H. Zhou and M. Li. Semi-supervised regression with co-training style algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 19(11):1479–1493, 2007.
57. Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.
58. Z.-H. Zhou and N. Li. Multi-information ensemble diversity. In *Proceedings of the 9th International Workshop on Multiple Classifier Systems*, pages 134–144, Cairo, Egypt, 2010.
59. Z.-H. Zhou, D.-C. Zhan, and Q. Yang. Semi-supervised learning with very few labeled training examples. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 675–680, Vancouver, Canada, 2007.
60. X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.
61. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, Washington, DC,

2003.